



دانشگاه پیام نور تهران شمال

دانشکده فنی و مهندسی

پایان نامه

کارشناسی ارشد - مهندسی کامپیوتر (نرم افزار)

گروه مهندسی کامپیوتر و فناوری اطلاعات

موضوع:

پیش بینی ترافیک با استفاده از الگوریتمهای تکاملی و داده های بزرگ

شهرز بذر افشان انجیلانی

استاد راهنما:

دکتر سید علی رضوی ابراهیمی

بهمن ۱۳۹۸

بسم الله الرحمن الرحيم

چکیده

هوشمند سازی یک شهر کاری مبتکرانه است که در بسیاری از شهرهای اروپایی این کار انجام شده است. از اهداف شهرهای هوشمند، بهبود برنامه ریزی، مدیریت اجتماعی و مدیریت زیربنایی شهر را می توان نام برد. شهر هایی چون دوبلین، لیون، آمستردام و بارسلونا نمونه هایی از شهرهای هوشمند هستند. بسیاری از شهرهای اروپایی از جمله دوبلین یک برنامه اطلاعاتی فعال دارند که با حفظ مسائل مربوط به حریم خصوصی بسیاری از پورتال های داده ای را باز می گذارند که در دسترس عموم قرار گیرند در چنین شهرهایی تجزیه و تحلیل ترافیک، و ارائه یک داشبورد بصری برای تحلیل آن امری مهم تلقی می شود. به دلیل اینکه روشهای سنتی قادر به پیشبینی با کیفیت مناسب نیستند تحقیق برای شناسایی روش های سنتی پیش بینی و تحلیل ترافیک در پیش بینی منطقه و کاوش داده های فضایی انجام می شود. محققان تجزیه و تحلیل داده ها را دلیل این کار برای شناسایی مشکلات و توسعه آنها می دانند. تحقیق بیشتر به شکل مفهوم چگونگی رویکرد تجزیه و تحلیل روش پیش بینی ترافیکی سنتی با داده های تاریخی و رسانه های اجتماعی خواهد بود. روش داده کاوی CRISP DM در طول پروژه مورد استفاده قرار خواهد گرفت. / این روش، یک استاندارد صنعتی برای استخراج داده است. این روش نقش مهمی در ایجاد تکنیک ها و ابزارها دارد. درکل روش ما مبتنی بر استفاده از داده هایی هست که در دیتاست استاندارد وجود دارند و بعد از پیش پردازش و اعمال الگوریتم نتایج آن بصورت تحلیلی ذکر خواهند شد

کلمات کلیدی: پیش بینی ترافیک، تجزیه و تحلیل الگوهای ترافیکی، داده کاوی

فهرست مطالب

فصل اول.....	۱
معرفی.....	۱
مقدمه.....	۲
۱-۱ تعریف مسئله و بیان سؤال‌های اصلی تحقیق.....	۲
۲-۱ اهمیت و ضرورت انجام تحقیق.....	۳
۳-۱ کاربردهای متصور از تحقیق.....	۴
۴-۱ پیشینه داده‌های بزرگ.....	۵
۵-۱ فرضیه‌ها.....	۶
۶-۱ اهداف.....	۶
۷-۱ استفاده کنندگان از تحقیق.....	۷
۸-۱ جنبه جدید و نوآوری طرح.....	۷
۹-۱ روش انجام تحقیق.....	۷
فصل دوم.....	۸
معرفی مفاهیم عمومی.....	۸
مقدمه.....	۹
۱-۲ داده، اطلاعات و دانش.....	۱۰
۲-۲ داده‌کاوی.....	۱۰
۳-۲ تعاریف داده‌کاوی.....	۱۱
۴-۲ ضرورت داده‌کاوی.....	۱۳
۵-۲ مراحل داده‌کاوی.....	۱۴

۱۴آماده‌سازی داده	۲-۵-۱
۱۵یادگیری مدل	۲-۵-۲
۱۵ارزیابی و تفسیر مدل	۲-۵-۳
۱۶روشهای یادگیری مدل در داده‌کاوی	۲-۶
۱۷مباحث داده‌کاوی	۲-۷
۱۸وظایف داده‌کاوی	۲-۸
۲۰NoSQL	۲-۹
۲۰Map Reduction	۲-۱۰
۲۲فصل سوم	
۲۳مقدمه	
۲۳۱-۳ کارهای پیشین و ادبیات تحقیق	
۲۵۱-۱-۳ سری زمانی پیش بینی	
۲۸۲-۱-۳ اثرات آب و هوا در ترافیک	
۳۰۳-۱-۳ تکنیک های فضایی	
۳۰۴-۱-۳ رسانه های اجتماعی	
۳۲۲-۳ کارهای داخلی	
۳۳۳-۳ داده های بزرگ	
۳۵۴-۳ کاهش نقشه	
۳۷۵-۳ داشبورد تحلیلی	
۴۰۶-۳ الگوریتم ها	
۴۲۷-۳ نتیجه گیری	
۴۳فصل چهارم	

۴۴	۱-۴ مجموعه داده های ترافیک
۴۵	۱-۱-۴ اطلاعات ترافیکی
۴۵	۲-۱-۴ داده اتصال
۴۶	۳-۱-۴ داده های مسیرها
۴۶	۲-۴ اطلاعات آب و هوا
۴۸	۳-۴ داده های توییت
۴۸	۴-۴ داده های ترافیک جدول زمانی کاربر توییت
۴۹	۵-۴ جمع آوری داده ها
۵۰	۶-۴ اکتشاف داده ها
۵۱	۱-۶-۴ بررسی ترافیک
۵۳	فصل پنجم
۵۴	۱-۵ انتخاب مدل زمان سفر استاندارد (STT)
۵۵	۲-۵ مدل آب و هوا
۵۹	۳-۵ اتصالات مدل پیش بینی
۶۰	۱-۳-۵ مجموعه داده های پیش بینی شده
۶۰	۲-۳-۵ نتایج پیش بینی
۶۲	۴-۵ مدل سازی ترافیک توییت
۶۳	فصل ششم
۶۴	۱-۶ پیاده سازی
۶۶	۲-۶ تحلیل ترافیک
۶۶	۳-۶ مدل پیش بینی
۶۷	۴-۶ داشبورد تجزیه و تحلیل

۶-۵ کار آینده..... ۶۷

مراجع..... ۶۸

واژه نامه..... ۷۴

فهرست جداول

جدول ۱-۱: مازول های پایتون اصلی	۴
جدول ۱-۳: مقایسه عملکرد پیش بینی کننده های مختلف	۲۷
جدول ۲-۳: خلاصه ای از تغییرات ARIMA [۲]	۲۷
جدول ۳-۳: مقایسه عملکرد با استفاده از RMSE ARIMA [۲]	۲۸
جدول ۴-۳: مقایسه شرایط مرطوب و خشک در جریان ترافیک	۲۹
جدول ۵-۳: بخشی از نام گفتار	۳۱
جدول ۱-۴: متغیرهای ثبت شده در داده های ترافیکی	۴۵
جدول ۲-۴: متغیرهای جدول اتصال	۴۶
جدول ۳-۴: متغیرهای داده های مسیر	۴۶
جدول ۴-۴: متغیرهای اطلاعات آب و هوایی	۴۷
در جدول ۵-۴: توزیع مقادیر از ۲۰۱۲/۰۷/۲۳ تا ۲۰۱۴/۰۴/۱۹ ۲۳:۵۰ نشان داده شده است	۵۱
جدول ۶-۴: توزیع مقادیر	۵۲
جدول ۱-۵: نواسانات ترافیکی	۵۴
جدول ۲-۵: همبستگی بارندگی و دما با ترافیک	۵۹
جدول ۳-۵: نتایج توییت	۶۲
جدول ۱-۶:	۶۴
جدول ۲-۶:	۶۵

فهرست اشکال

- شکل ۱-۱: رویکرد تقسیم داخلی و تسخیر داخلی MongoDB ۶
- شکل ۱-۳: مقایسه جریان پیش بینی شده سفر با مدل های مختلف ۲۶
- شکل ۲-۳: بخشی از الگوی حکم سخنرانی ۳۱
- شکل ۳-۴: نقشه و کاهش [۱] ۳۵
- شکل ۳-۵: پیش بینی برخورد [۱] ۳۶
- شکل ۳-۶: تجسم رنگ داده های بزرگ ۳۸
- شکل ۳-۸: الگوریتم های Sci-Py [۲۶] ۴۱
- شکل ۴-۱: نقشه گوگل دوبلین ۴۵
- شکل ۴-۲: مکانهای ایستگاه هواشناسی ۴۷
- شکل ۴-۳: نمونه ای از توییتها ۴۹
- شکل ۴-۴: نحوه جمع آوری داده ها ۵۰
- شکل ۴-۵: مشاهدات زمان سفر ۵۲
- شکل ۵-۱: نمودار همبستگی روزانه ۵۵
- شکل ۵-۲: همبستگی زمانهای اوج ترافیک و بارندگی ۵۶
- شکل ۵-۳: همبستگی زمانهای اوج ترافیک و بارندگی ۵۷
- شکل ۵-۴: همبستگی زمانهای اوج ترافیک و دما ۵۸
- شکل ۵-۵: نتایج رگرسیون خطی ۶۱
- شکل ۶-۵: نتایج رگرسیون برداری پشتیبان ۶۱

فصل اول

معرفی

مقدمه

هوشمند سازی شهر، کاری مبتکرانه محسوب می شود که در بسیاری از شهرهای اروپایی به انجام رسیده است. اهداف شهرهای هوشمند بهبود برنامه ریزی، مدیریت اجتماعی و مدیریت زیربنایی است. نمونه ای از این شهر هایی که هوشمند سازی در آنها انجام شده است عبارتند از دوبلین، لیون، آمستردام و بارسلونا.

بسیاری از شهرهای اروپایی از جمله دوبلین یک برنامه اطلاعاتی فعال دارند که با حفظ مسائل مربوط به حریم خصوصی بسیاری از پورتال های داده ای را باز می گذارند که در دسترس عموم قرار گیرند. شورای شهر دوبلین بیش از ۲۵۰ مجموعه داده موجود را برای شهروندان فراهم کرده است [۳]

شبکه های حسگر بی سیم تکنولوژی ای است که نقش مهمی در شهرهای هوشمند ایفا می کند. دوبلین همراه با بسیاری از شهرهای دیگر از این تکنولوژی برای جمع آوری اطلاعات مربوط به ترافیک استفاده می کنند. هدف این است که یک زیرساخت کامل داشته باشیم که نظارت بر رفتارهای ترافیکی را امکان پذیر می سازد تا تصمیمات در مورد توسعه شهرها بتواند به شیوه ای دقیق گرفته شود. در نظر گرفتن متغیرهایی مانند شرایط آب و هوایی و فصلی می تواند تصمیم به طراحی شبکه جاده ای را بهبود بخشد.

۱-۱ تعریف مسئله و بیان سؤالهای اصلی تحقیق

فناوری های نوین برای جمع آوری و ذخیره ای اطلاعات مقادیر عظیمی از داده ها را در دسترس حوزه های کاربردی مختلف از قبیل دنیای کسب و کار، بانکداری، امور پزشکی، علمی و غیره قرار داده اند. مجموعه فعالیت های انجام شده جهت تحلیل این پایگاه های داده بزرگ تحت عناوین مختلفی از قبیل داده کاوی^۱، اکتشاف دانش، شناسایی الگو و یادگیری ماشین بیان می شود که معمولاً باهدف استخراج دانش مفید برای پشتیبانی تصمیم صورت می گیرد.

^۱ Data Mining

در این پژوهش سعی می شود برای تحلیل از مجموع داده های سنجش از راه دور و رسانه های اجتماعی با منابع داده ای باز استفاده شود که به دلیل تنوع این منابع در این پژوهش ما را با ۴ نوع چالش های داده بزرگ^۲ مواجه می کند.

در این راستا سوال زیر مطرح می شود که در عمل به آن پاسخ داده می شود:

آیا می توانیم برای پیش بینی و تجزیه و تحلیل ترافیک، داده های رسانه های اجتماعی را به عنوان بخشی از داده های یک شهر، مورد استفاده قرار دهیم؟

۱-۲ اهمیت و ضرورت انجام تحقیق

امروزه داده ها قلب تپنده فرآیند تجاری بیشتر سازمان ها تلقی می شوند، لذا نیاز به ابزاری است که بتوان داده های ذخیره شده را پردازش کرده و اطلاعات حاصل از این پردازش را در اختیار سازمان ها قرار دهد. در این راستا دانش داده کاوی دانشی است که در سال های اخیر گسترش فوق العاده سریعی در دنیا داشته است. انگیزه برای گسترش داده کاوی به طور عمده از دنیای تجارت در دهه ۱۹۹۰ پدید آمد. داده کاوی فرآیند کشف دانش پنهان درون داده ها است که با توصیف، تشریح، پیش بینی و کنترل پدیده های گوناگون پیرامونی، دارای کاربرد بسیار وسیعی در حوزه های مختلف است به گونه ای که مرز و محدودیتی برای کاربرد آن در نظر گرفته نشده و زمینه های کاربردی آن را از ذرات کف اقیانوس تا اعماق فضا می دانند (شهرابی، ۱۳۸۶).

تحقیقاتی برای شناسایی روش های سنتی پیش بینی و تحلیل ترافیک در پیش بینی منطقه و کاوش داده های فضایی انجام شده است. در این تحقیقات مشکلات در ارائه تجزیه و تحلیل توسط محققان شناسایی شده اند. تحقیق انجام شده بیشتر به شکل مفهوم چگونگی رویکرد به تجزیه و تحلیل روش پیش بینی ترافیکی سنتی با داده های تاریخی و رسانه های اجتماعی می پردازد

² Big Data 4 V

تجزیه و تحلیل ترافیک شهری، و ارائه یک داشبورد بصری برای تحلیل ترافیک برای یک شهر امری حیاتی است. تامین داده های اولیه تحلیل مورد نظرا از طریق داده های سنجش از راه دور با اطلاعات رسانه های اجتماعی از منابع داده ای باز، متفاوت است که در این پژوهش ما را با ۴ نوع چالش های داده بزرگ^۳ مواجه می کند. از این جهت در این پژوهش از روش داده کاوی CRISP DM استفاده شده است. CRISP DM یک استاندارد صنعتی برای استخراج داده است. این روش نقش مهمی در تکنیک ها و ابزارها دارد.

در این کار مواردی هست که نیاز به توجه بیشتری دارند. بسیاری از داده ها قبل از پردازش، پردازش زبان طبیعی، الگوریتم های ریاضی، ادغام توئیتر، کاوش داده های فضایی، تجسم، برنامه وب، ذخیره سازی داده های ساختار یافته و غیر ساختاری وجود دارد. این ابزارها در متلب و پایتون موجودند.

<i>Package</i>	<i>Description</i>
SciPy	Scientific Algorithms and Methods
NumPy	Number Manipulation
Django	Web Application
TwitterAPI	Twitter Integration
PyMongo	MongoDB NoSQL Integration

جدول ۱-۱: ماژول های پایتون اصلی

۳-۱ کاربردهای متصور از تحقیق

داده کاوی روشی است که بامطالعه و بررسی الگوریتم های مؤثر و کارآمد جهت تبدیل مقادیر زیاد داده به معلومات مفید می پردازد. این همان دلیل اصلی موردتوجه قرار گرفتن داده کاوی است. باگذشت زمان و با برطرف شدن نیازهای اولیه در داده کاوی توسط الگوریتم های پایه ای ارائه شده، کم کم تفکر ایجاد کیفیت و سرعت بهتر و بالاتر به وجود آمد. با بروز این احساس نیاز، محققان سعی در بهینه نمودن الگوریتم های مختلف داده کاوی نمودند. با توجه به اینکه تحلیل ترافیک کمک بسیاری در جلوگیری از هدر رفت زمان

^۳ Big Data 4 V

دارد، از اهمیت زیادی برخوردار است. در این پژوهش سعی می‌شود با ترکیب اطلاعات مربوط به آب و هوا، ترافیک و رسانه‌های اجتماعی تحلیل جدید از ترافیک به عمل آید که به مسئولان درون شهری و ترافیک برای برنامه ریزی در جهت جلوگیری از هدر رفت زمان و سوخت کمک کند..

۱-۴ پیشینه داده های بزرگ^۴

در سراسر تکنیک های پایان نامه که در تمام مراحل توسعه برای رسیدگی به چهار چالش از داده های بزرگ: حجم، سرعت، تنوع و حقیقت استفاده می شود.

- حجم داده های ترافیکی چالشی است که با استفاده از MapReduce کاهش می یابد با دسته بندی داده های مرتبط با هم که به سیستم پایگاه داده اجازه می دهد تا جستجویی به طور موثر از طریق مکانیزم فهرست بندی انجام دهیم.

- سرعت: داده های توییت شده را برای این سیستم در زمان واقعی بتوان خواند. برای پردازش و ذخیره داده ها مجدداً MapReduce و Indexing استفاده می شود.

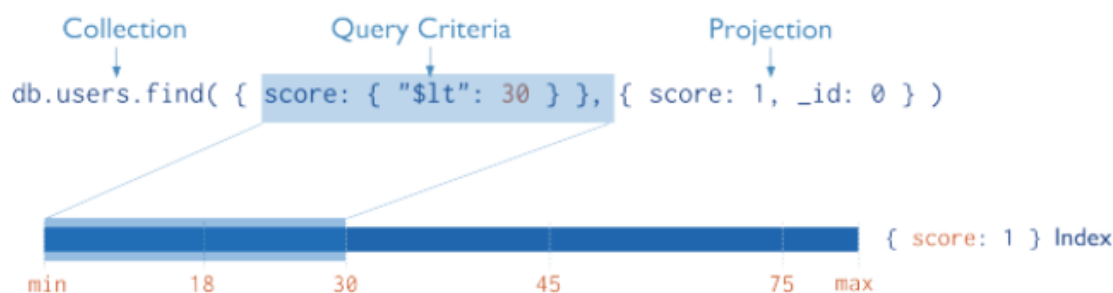
- انواع: ترافیک منابع داده، آب و هوا و توییت حاوی انواع داده هایی مانند نشانه های زمانی، جغرافیایی، رشته ها و عدد صحیح است. سیستم پایگاه داده به نام MongoDB برای این نیازها مورد استفاده قرار میگیرد.

- اطمینان در این مورد ذخیره داده ها در تجزیه و تحلیل برای استفاده در تجزیه و تحلیل های بعدی است.

تکنیک برای غلبه بر سیستم های داده بزرگ تقسیم تسخیر MongoDB NoSQL است [۴] که چالش های ۴V را [۵] در نظر می گیرد و به کاربران اجازه می دهد یک طراحی را به گونه ای انجام دهند که داده ها بتوانند به طور قابل توجهی ذخیره و بازیابی شوند .

⁴ Big Data

NoSQL پایگاه داده بزرگ اطلاعاتی سبک وزن است. هنگام اجرای یک راه حل، طراحی پایگاه داده، ایجاد شاخص ها بسیار مهم است. این شاخص ها اجازه می دهد که سیستم مجموعه ای را به بخش تقسیم کند. در پس زمینه، پایگاه داده جمع آوری شده است و سپس یک جدول فهرست برای نقشه برداری داده ها و موقعیت آن در سیستم فایل ایجاد می شود. به عنوان مثال در شکل ۱-۱ یک مجموعه کاربر در سیستم فایل براساس نمره معیارهای index chunked شده است. این منجر به بازیابی سریع مجموعه ها می شود زیرا پایگاه داده می داند فقط فایل هایی را که حاوی اطلاعات مرتبط هستند را جستجو می کند.



شکل ۱-۱ رویکرد تقسیم داخلی و تسخیر داخلی MongoDB

برای جمع آوری داده های سری زمانی، داده ها براساس زمان بندی و یا مکانی نشان داده می شوند.

۵-۱ فرضیه ها

در این پژوهش فرضیه مطرح به شرح زیر می باشد:

آیا می توانیم از داده های دردسترس و رسانه های اجتماعی برای پیش بینی و تجزیه و تحلیل ترافیک به عنوان بخشی از اطلاعات یک شهر استفاده کنیم؟

۶-۱ اهداف

اهداف اصلی:

- به دست آوردن مدل پیش بینی عمومی با استفاده از ذخیره تاریخ ترافیکی و اطلاعات مربوط به آب و هوا

- طراحی روشی برای ارائه تحلیل رویدادهای مرتبط با ترافیک با ذخیره داده های توییتري

اهداف فرعی:

- ایجاد یک داشبورد تجزیه و تحلیل برای الگوهای ترافیکی با استفاده از تکنیک های داده کاوی مدل های پیش بینی و تجزیه و تحلیل توییتري.

۷-۱ استفاده کنندگان از تحقیق

- سازمان اطلاع و کنترل ترافیک
- محققان داده کاوی
- مرکز مدیریت راه های کشور
- شهرداری

۸-۱ جنبه جدید و نوآوری طرح

در این پایان نامه تجزیه و تحلیل ترافیک شهری انجام شده، و یک داشبورد بصری برای تحلیل ترافیک یک شهر ارائه می گردد. از طرفی مجموعه داده های اولیه با استفاده از داده های سنجش از راه دور و اطلاعات رسانه های اجتماعی از منابع داده های باز، که متفاوت است مورد بحث قرار گرفته است

۹-۱ روش انجام تحقیق

داده ها هم تفسیر آزمایشی دارند هم توصیفی. آزمایشی به این دلیل که قابل تست روی جامعه واقعی می باشند و ذینفعان آن کنترل ترافیک کشوری و در بعد عملی کوچک آن کنترل ترافیک شهری را می تواند مدل کرد. توصیفی هم به این دلیل که با روشهای دیگر قابل مقایسه خواهد بود تا دقت و سرعت الگوریتم مورد ارزیابی قرار گیرد. در کل روش ما مبتنی بر استفاده از داده هایی هست که در دیتاست استاندارد وجود دارند و بعد از پیش پردازش و اعمال الگوریتم نتایج آن بصورت تحلیلی ذکر خواهند شد

فصل دوم

معرفی مفاهیم عمومی

مقدمه

با گسترش فناوری اطلاعات و ارتباطات^۹ در جهان و ورود سریع آن به زندگی روزمره مردم، مسائل و ضرورت‌های تازه‌ای به‌جان آمده. امروزه انسان توسعه‌یافته کسی است که به اطلاعات دسترسی داشته باشد و دسترسی به اطلاعات یک قدرت محسوب می‌شود. در نتیجه تلاش برای استخراج اطلاعات از داده‌ها توجه بسیاری از افراد دخیل در صنعت اطلاعات و حوزه‌های وابسته را به خود جلب نموده است (Hand, 1998).

حجم بالای داده‌های دائماً در حال رشد در همه حوزه‌ها و نیز تنوع آن‌ها به شکل داده متنی، اعداد، نقشه‌ها، عکس‌ها، تصاویر ماهواره‌ای و عکس‌های گرفته‌شده با اشعه ایکس نمایانگر پیچیدگی کار تبدیل داده‌ها به اطلاعات است (Hand, 1998) با توجه به تنوع زیاد مخاطبین، مشتریان، بازارها، تنوع و پیچیدگی خدمات و محیط‌های کسب‌وکار و لزوم دسترسی به اطلاعات مناسب برای تصمیم‌گیری صحیح و به‌موقع، استفاده از راهکارهای مناسب برای طبقه‌بندی و یافتن اطلاعات کاربردی و اثربخش از میان انبوهی از داده‌ها برای سازمان‌ها امری ضروری و حیاتی بوده و یک تخصص و هنر محسوب می‌شود. استراتژی‌ها و فنون متعددی برای گردآوری، ذخیره، سازمان‌دهی و مدیریت کارآمد داده‌های موجود و رسیدن به نتایج معنی‌دار بکار گرفته‌شده‌اند. این تلاش‌ها را می‌تواند به‌عنوان یک حرکت پیش‌رونده از ایجاد یک بانک اطلاعات ساده تا شبکه‌ها و بانک‌های اطلاعاتی رابطه‌ای و سلسله‌مراتبی برای پاسخگویی به نیاز روزافزون سازمان‌دهی و بازیابی اطلاعات ملاحظه نمود. داده‌کاوی یکی از پیشرفت‌های اخیر در راستای فناوری‌های مدیریت داده‌هاست که در واقع پاسخی به این نیاز سازمان‌ها و مؤسسات است. داده‌کاوی مجموعه‌ای از فنون است که به شخص امکان می‌دهد تا ویرای داده‌پردازی معمولی حرکت کند و به استخراج اطلاعاتی که در انبوه داده‌ها مخفی و یا پنهان است کمک می‌کند. هر چه حجم داده‌ها بیشتر

^۹ Information and Communication Technology (ICT)

و روابط میان آن‌ها پیچیده‌تر باشد، دسترسی به اطلاعات نهفته در داده‌ها شده لذا نقش داده‌کاوی به‌عنوان یکی از روش‌های کشف دانش، روشن‌تر می‌شود (شهرابی، ۱۳۸۶).

در این فصل به معرفی مفاهیم عمومی مورد نیاز که در ادامه پژوهش مورد استفاده قرار خواهند گرفت می‌پردازیم.

۲-۱ داده، اطلاعات و دانش

ابتدا پیش از هر چیز بهتر است به بررسی مفاهیم داده، اطلاعات و دانش بپردازیم.

داده تنها مقادیری خام است. این مقادیر در رابطه با اشیاء، رخدادها، فعالیت‌ها و تراکنش‌هایی است که تولید، طبقه‌بندی و ذخیره‌شده است؛ ولی به‌گونه‌ای سازمان‌یافته نیست که بتوان از آن‌ها معنا و مفهوم خاصی را استنباط نمود. این داده‌ها می‌تواند به شکل عددی، متنی، نمودار، صدا، تصویر و غیره باشد (جام سحر، ۱۳۸۹).

چنانچه از مجموعه‌ای از داده‌ها که به شکل خاصی سازماندهی شده‌اند بتوان معنا و مفهومی را استنباط و یا درک کنیم، به این دریافت «اطلاعات» می‌گوییم. در حقیقت اطلاعات به نحوه‌ی تفسیر کردن داده‌ها و معنایی که از آن دریافت شده است اطلاق می‌گردد. اگر داده‌ها و یا اطلاعات به شکلی سازمان‌دهی شوند که منجر به حل مسئله‌ای شود و یا سبب گردد تصمیمی اتخاذ گردد به چنین درکی «دانش» می‌گویند (جام سحر، ۱۳۸۹).

۲-۲ داده‌کاوی

فناوری‌های نوین اطلاعاتی و ارتباطی و همچنین فناوری‌های پشتیبان تصمیم، با جمع‌آوری، ذخیره، ارزیابی، تفسیر و تحلیل، بازیابی و اشاعه اطلاعات و دانش به کاربران خاص، می‌توانند در اطلاع‌یابی به‌موقع، صحیح و موردنیاز به افراد تأثیر زیادی داشته باشند. یکی از ابزارهای مورد استفاده در این فناوری‌ها، داده‌کاوی می‌باشد. داده‌کاوی شامل استفاده از ابزارهای پیشرفته تحلیل داده به‌منظور کشف الگوهای معتبر ناشناخته و روابط در مجموعه داده‌های حجیم است (Hand, 1998).

تاریخچه‌ی کشف دانش از پایگاه‌های اطلاعاتی قدمت چندانی ندارد و امروزه به داده‌کاوی مشهور است. اصطلاح کشف دانش نخستین بار در دهه‌ی ۱۹۹۰ مطرح شد و توجه پژوهشگران را به الگوریتم‌های داده‌کاوی معطوف کرد. نام دیگر داده‌کاوی، کشف دانش در پایگاه داده یا به اختصار ^۱KDD است.

علم داده‌کاوی از علوم مختلفی از جمله علم آمار، هوش مصنوعی، یادگیری ماشین، شناسایی الگو و پایگاه داده نشأت گرفته است. درواقع این علوم ریشه‌های علم داده‌کاوی هستند. می‌تواند از تمام فنون موجود در این علوم برای درک چگونگی کارکرد الگوریتم‌ها و روش‌های داده‌کاوی استفاده نمود. الگوریتم‌ها موجود در هوش مصنوعی و علم آمار کمک زیادی به داده‌کاوی می‌کنند. مباحث موجود در یادگیری ماشین و شناسایی الگو نیز با مباحثی که در داده‌کاوی هستند همپوشانی دارند. در علم پایگاه داده یک پایگاه داده بزرگ داریم که آن را به عنوان انبار داده می‌شناسیم. این انبار داده باید حتماً وجود داشته باشد با یک الگوریتم داده‌کاوی بتواند روی آن کار کند. حال از فنی که در پایگاه داده برای جمع‌آوری داده وجود دارد می‌توانیم برای ایجاد یک انبار داده جهت استفاده در فرآیند داده‌کاوی استفاده کنیم. داده‌کاوی فراتر از جمع‌آوری و مدیریت داده است و شامل تجزیه و تحلیل و پیشگویی می‌شود (صنّعی آبادی و دیگران، ۱۳۹۳).

۲-۳ تعاریف داده‌کاوی

تعاریف گوناگونی برای داده‌کاوی ارائه شده است. در برخی از این تعاریف داده‌کاوی در حد ابزاری که کاربران را قادر به ارتباط مستقیم با حجم عظیم داده‌ها می‌سازد معرفی گردیده است و در برخی دیگر، تعاریف دقیق‌تر که در آن‌ها به کاوش در داده‌ها توجه می‌شود موجود است. برخی از این تعاریف عبارت‌اند از:

^۱ Knowledge Discover in Database

▪ داده کاوی عبارت است از فرایند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه داده‌های بزرگ که از آن جهت تصمیم‌گیری در فعالیتهای تجاری استفاده می‌شود.

▪ اصطلاح داده کاوی به فرایند نیم خودکار تجزیه و تحلیل پایگاه داده‌های بزرگ به منظور یافتن الگوهای مفید اطلاق می‌شود.

▪ داده کاوی به معنی جستجو در یک پایگاه داده برای یافتن الگوهایی میان داده‌ها است.

▪ داده کاوی یعنی تجزیه و تحلیل مجموعه داده‌های قابل مشاهده برای یافتن روابط مطمئن بین داده‌ها.

▪ داده کاوی، استخراج یا اقتباس دانش از مجموعه‌ی داده‌ها است و به فرایندی گفته می‌شود که دانش را از داده‌ها استخراج می‌کند و این دانش در قالب الگوها و مدل‌ها بیان می‌شود (تقوی فرد و نادعلی، ۱۳۹۱).

▪ داده کاوی به بررسی و تجزیه و تحلیل مقادیر عظیمی از داده‌ها به منظور کشف الگوها و قوانین معنی دار اطلاق می‌شود که عمدتاً از طریق ساختن مدل‌ها و الگوریتم‌ها، ورودی‌ها را باهدف یا مقصد خاصی مرتبط می‌نماید (شهرابی، ۱۳۸۶).

همان‌گونه که در تعاریف گوناگون داده کاوی مشاهده می‌شود، به مفاهیمی چون استخراج دانش، تحلیل و یافتن الگوی بین داده‌ها اشاره شده است. با توجه به تعاریف مختلفی که برای داده کاوی وجود دارد شاید بتوان تعریف زیر را به عنوان تعریفی جامع برای داده کاوی به کار برد. «استخراج خودکار دانش جدید و مفید از منابع داده‌ای حجیم موجود طی یک فرآیند غیر بدیهی مشخص، داده کاوی نامیده می‌شود»

هدف اصلی داده کاوی کشف دانش است. این دانش نظامی خواهد بود که در داده‌ها وجود دارد. پس از کشف دانش ممکن است دو حالت وجود داشته باشد: حالت اول آن است که افراد خبره در دامنه داده‌ی مورد کاوش، آگاه به دانش استخراج شده باشند. در این صورت آن دانش به عنوان یک قانون صحیح تلقی خواهد شد. در حالت دوم ممکن است دانش کشف شده، یک دانش جدید بوده و در بین افراد متخصص

در آن زمینه شناخته شده نباشد. حال این دانش بررسی شده و در صورت منطقی بودن تبدیل به فرضیه شده و در نهایت در ست یا غلط بودن این فرضیه با آزمایشان و بررسی‌های متعدد اثبات می‌شود. این فرضیه در صورت درست بودن به قانون تبدیل خواهد شد (صنّعی آبادی و دیگران، ۱۳۹۳).

۲-۴ ضرورت داده‌کاوی

روش‌های بهبود مدیریت داده نقش مهمی در افزایش قابلیت استفاده از اطلاعات و کاهش هزینه‌های ذخیره‌سازی دارند. زیرا طی دهه‌های گذشته شاهد افزایشی سریع در حجم اطلاعات جمع‌آوری و ذخیره‌شده (با نظر گرفتن این فرض که هر سال تقریباً حجم داده جهان دو برابر می‌شود) هستیم. با نظر به افزایش سرعت قدرت رایانه‌ها در دهه‌های گذشته و افزایش تعداد مجموعه‌های بزرگ داده و شناخت ارزش تغییرات ملایم^۷ دیگر، روش‌های سنتی به‌تنهایی قادر به ارائه تحلیل‌هایی قدرتمند از داده نیستند و این خود بر ضرورت استفاده از قابلیت بالای روش‌شناسی تحلیل رایانه تأکید می‌کند (Dillon, 1998).

بسیاری از سازمان‌ها چه در بخش خصوصی و چه در بخش دولتی، با استفاده از فناوری‌ها و فرایندهای کسب‌وکار پیشرو، جذب داده‌کاوی شده‌اند. بعضی از این تغییرات عبارت‌اند از: رشد شبکه‌های رایانه‌ای، توسعه فنون جستجو مانند شبکه‌های عصبی و الگوریتم‌های پیشرفته، گسترش مدل‌های مشتری - کارگر، افزایش میزان دسترسی کاربران به منابع مرکزی داده و افزایش توانایی ترکیب داده از منابع گوناگون و تبدیل آن به یک منبع واحد قابل جستجو. به‌علاوه، سازمان‌ها از داده‌کاوی به‌عنوان ابزاری در تحقیقات مشتری، صرفه‌جویی و تحقیقات پزشکی نیز استفاده می‌کنند.

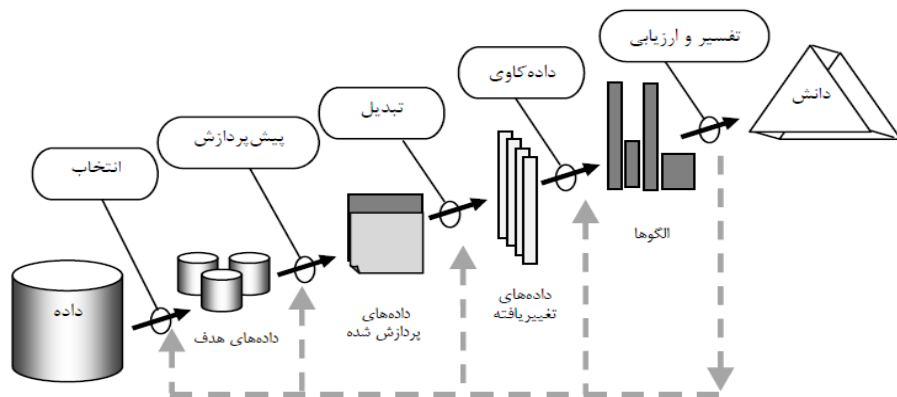
داده‌کاوی از دو گام اصلی پیش‌پردازش داده و شناخت الگو تشکیل می‌شود. پردازش داده شامل مواردی است که یا تعداد ترکیب‌های عناصر داده زیاد باشد و یا نشانه‌ها، برگرفته از چندین داده‌ی ساده باشند. معمولاً فرایند پیش‌پردازش زمان‌بر است. شناخت الگو نیز در جایگاه الگوی داده ترکیبی باشد، استفاده می‌شود (Dillon, 1998).

⁷ Untapped

۵-۲ مراحل داده‌کاوی

فرایند داده‌کاوی شامل سه مرحله است: آماده‌سازی داده، یادگیری مدل، ارزیابی و تفسیر مدل.

شکل ۱-۲ این مراحل سه‌گانه را به همراه زیر مراحل و همچنین ورودی/خروجی‌های آن‌ها نشان می‌دهد. در ادامه به توصیف هر کدام از این مراحل خواهیم پرداخت (صنّعی آبادی و دیگران، ۱۳۹۳).



شکل ۱-۲ مراحل داده‌کاوی (صنّعی آبادی و دیگران، ۱۳۹۳).

۱-۵-۲ آماده‌سازی داده

اولین و مهم‌ترین مرحله در فرایند داده‌کاوی آماده‌سازی داده است. هدف در این مرحله تأمین ورودی مناسب برای مرحله‌ی حیاتی یادگیری مدل است. در این مرحله داده پردازش نشده از کل منابع داده‌ای موجود (که ممکن است توزیع شده نیز باشند) استخراج شده، سپس در مرحله‌ای مستقل مورد پردازش اولیه قرار می‌گیرد. خروجی در مرحله آماده سازی داده عبارت است از داده پیش‌پردازش شده که امکان یادگیری مدل از روی آن وجود دارد.

همان‌گونه که بیان شد اولین گام در مرحله‌ی آماده‌سازی داده، استخراج داده از منابع داده‌ای موجود می‌باشد. در این گام می‌بایست داده‌ها از منابع مختلفی که پراکنده‌اند، به‌صورت متمرکز در یک محل جمع‌آوری شده و یک انبار داده مرکزی ایجاد شود. دلیل اصلی این گردآوری آن است که در اغلب موارد

داده به صورت متمرکز در یک مکان وجود ندارد. به علاوه داده‌ها در بخش‌های مختلف ممکن است در فرمت‌های گوناگونی نیز ذخیره شده باشند. مثلاً در پایگاه‌های داده‌ای و یا فایل‌های اکسل یا متنی قرار گرفته باشند. دومین گام در مرحله‌ی آماده‌سازی داده، پیش‌پردازش داده‌های استخراج شده است. مهم‌ترین رسالت این گام زدودن مشکلات مختلفی است که در داده‌ها وجود دارند. این مشکلات در داده‌ها مانع از آن می‌شوند که مرحله‌ی یادگیری مدل بتواند نظم واقعی را در داده بیابد. در هر حال پس از پایان مرحله‌ی آماده‌سازی داده، مجموعه داده‌ای آماده خواهد شد که فاقد مشکلات جدی و کلیدی است و امکان کشف دانش نهفته در آن با استفاده از مرحله‌ی یادگیری مدل وجود دارد.

۲-۵-۲ یادگیری مدل

در این مرحله با استفاده از الگوریتم‌های متنوع و با توجه به ماهیت داده، سعی بر این است که نظم‌های مختلف موجود در داده را شناسایی نموده و در فرمتی مشخص به عنوان دانش نهفته در داده ارائه کنیم. برای یادگیری مدل می‌بایست روش‌های آن را به درستی شناخت تا بتوان در جای مناسب، روش درست را انتخاب نمود و به کار بست.

۲-۵-۳ ارزیابی و تفسیر مدل

در این مرحله دانش تولید شده در مرحله قبل ارزیابی شده و مورد تفسیر قرار می‌گیرد. منظور از ارزیابی دانش آن است که می‌بایست میزان صحت دانش تولید شده مشخص شود تا بتوان به آن اعتماد نمود و به صورت عملی از آن استفاده کرد. روش‌های مختلفی برای ارزیابی دانش تولید شده وجود دارند که رابطه‌ی تنگاتنگی با روش یادگیری مدل دارند.

تفسیر مدل به معنای آن است که دانش تولید شده را مورد بررسی قرار داده و توجیهی معنایی جهت تبیین منطق آن ارائه نماییم. در صورت قابل تفسیر بودن دانش تولید شده، انجام این کار بسیار ساده است (به عنوان مثال زمانی که دانش به صورت درخت و یا مجموعه قوانین باشد). در مقابل امکان تفسیر مدل برای مواقعی که دانش به صورت غیر قابل تفسیر باشد (مانند دانش تولید شده توسط شبکه‌های عصبی و یا ماشین بردار پشتیبان) بسیار مشکل‌تر و شاید غیرممکن خواهد بود.

۶-۲ روش‌های یادگیری مدل در داده‌کاوی

روش‌های مختلف داده‌کاوی را می‌تواند به دو گروه روش‌های توصیفی و روش‌های پیش‌بینی طبقه‌بندی نمود.

هدف از به کارگیری فنون پیش‌بینی کننده، پیش‌بینی ارزش یک ویژگی خاص بر اساس سایر ویژگی‌هاست. ویژگی مورد پیش‌بینی هدف نامیده شده و وابسته به سایر ویژگی‌هاست و ویژگی‌هایی که کمک به پیش‌بینی می‌کنند متغیرهای توضیحی و مستقل هستند (Tan, Steinbach, Kumar, 2005). در متون علمی مختلف روش‌های پیش‌بینی با نام روش‌های با ناظر^۸ نیز شناخته می‌شوند. روش‌های دسته‌بندی^۹، رگرسیون^{۱۰} و تشخیص انحراف^{۱۱} سه روش یادگیری مدل در داده‌کاوی با ماهیت پیش‌بینی هستند (صنّعی آبادی و دیگران، ۱۳۹۳).

اما هدف از به کارگیری فنون توصیفی استخراج الگوست به نحوی که ارتباط بین لایه‌های زیرین داده‌ها را خلاصه سازی کند. این روش‌ها الگوهای قابل توصیفی را پیدا می‌کنند که روابط حاکم بر داده‌ها را بدون در نظر گرفتن هرگونه برچسب و یا متغیر خروجی تبیین نمایند. در متون علمی مختلف روش‌های توصیفی با نام روش‌های بدون ناظر نیز شناخته می‌شوند. روش‌های خوشه‌بندی، کاوش قوانین انجمنی^{۱۲} و کشف الگوهای ترتیبی^{۱۳} مواردی از این دست هستند (صنّعی آبادی و دیگران، ۱۳۹۳).

^۸ Supervised Methods

^۹ Classification

^{۱۰} Regression

^{۱۱} Anomaly Detection

^{۱۲} Association Rule Mining

^{۱۳} Sequential Pattern Discovery

۷-۲ مباحث داده‌کاوی

علاوه بر اهمیت قابلیت‌های مربوط به شیوه فرایند داده‌کاوی در کشف و تحلیل داده، عوامل دیگری نیز مانند کیفیت داده، توانایی عملکرد^{۱۴}، هدفمندی^{۱۵}، محدودیت دسترسی عمومی و ... بر موفقیت نتایج طرح تأثیر می‌گذارند (Seifert, 2004)

■ کیفیت داده

کیفیت داده مبحثی چالش‌برانگیز در فرایند داده‌کاوی است و به صحت و کامل بودن داده برمی‌گردد. بعلاوه، داده‌کاوی می‌تواند بر ساختار و درجه سازگاری داده تحلیل‌شده نیز اثر بگذارد. تکرار و نسخه‌برداری از داده‌های ضبط‌شده، عدم وجود استانداردها، شرایط زمانی در به‌روزرسانی و خطاهای انسانی از جمله عوامل اثرگذار بر اثربخشی فنون پیچیده داده‌کاوی هستند (Seifert, 2004).

■ توانایی عملکرد

با توجه به کیفیت داده، مبحث توانایی عملکرد به پایگاه‌های متفاوت داده و نرم‌افزارهای داده‌کاوی برمی‌گردد. این مفهوم توانایی سامانه رایانه‌ای و یا داده را برای کار با سایر سامانه‌ها یا استفاده از داده در فرایندهای جاری و مرسوم، تعیین می‌کند.

توانایی عملکرد پایگاه‌های داده و نرم‌افزار مورد استفاده، در فرایند داده‌کاوی نقشی مهم دارند. زیرا در قابلیت جستجو و تحلیل همزمان پایگاه‌های چندگانه داده و تضمین سازش‌پذیری فعالی ته‌ای فرایند داده‌کاوی مؤثر هستند.

■ هدفمندی

¹⁴ Interoperability

¹⁵ Mission Creep

هدفمندی، یکی از مخاطره‌های مشخص داده‌کاوی است و بیانگر نحوه کنترل اطلاعات فرد است. این مفهوم توجه کاربر را به هدف داده‌کاوی و اولویت جمع‌آوری داده جلب می‌کند. یکی از دلایل اولیه اشتباه در نتایج، وجود داده‌های نادرست^{۱۶} است. البته اگر خود داده ارزش اقتصادی بالایی نداشته باشد، هزینه کسب اطمینان از صحت داده هرگز توجیه‌پذیر نخواهد بود.

■ محدودیت دسترسی عمومی^{۱۷}

از عامل اثرگذار دیگر، امنیت داده و محدودیت دسترسی به آن است. این مفهوم نقشی مهم در میزان تسهیم اطلاعات و شروع فرایند داده‌کاوی خواهد داشت (Seifert, ۲۰۰۴).

۲-۸ وظایف داده‌کاوی

به‌طور کلی اگر وظایف اصلی داده‌کاوی را در طبقه‌بندی، انتخاب و استخراج بدانیم هر کدام از این وظایف می‌توانند به‌منزله مشکلی در نظر گرفته شوند که الگوریتم داده‌کاوی به رفع آن می‌پردازد.

■ طبقه‌بندی

طبقه‌بندی وظیفه‌ای قابل‌توجه است. این وظیفه طی دهه‌های گذشته از طریق یادگیری ماشین و جوامع آماری مطالعه شده است. هدف این وظیفه، پیش‌بینی ارزش صفت موردنظر کاربر، مبتنی بر ارزش سایر صفتهایی (خصایصی) است که صفات پیش‌بینی‌کننده نام دارند.

قوانین طبقه‌بندی که به‌صورت اگر - آنگاه هستند، نوع خاصی از قوانین پیش‌بینی در نظر گرفته می‌شوند. قسمت اگر در این قوانین، شامل ترکیبی از موقعیت‌های ارزش صفت پیش‌بینی است و قانون نتیجه در قسمت آنگاه، شامل ارزش پیش‌بینی‌شده برای صفت هدف است.

¹⁶ Inaccurate

¹⁷ Privacy

در وظیفه طبقه‌بندی، داده‌ی کاویده شده به دو مجموعه آموزشی و آزمون تقسیم می‌شود. الگوریتم داده‌کاوی تنها با دسترسی به مجموعه آموزش به کشف قوانین می‌پردازد. برای این منظور، الگوریتم موردنظر باید هم به ارزش صفت‌های پیش‌بینی و هم به صفت هدف هر نگاشته (مدرک) در مجموعه آموزش دسترسی داشته باشد. زمانی که فرایند طبقه‌بندی به پایان می‌رسد و الگوریتم، مجموعه‌ای از قوانین را پیدا می‌کند؛ قابلیت پیش‌بینی این قوانین در مجموعه آزمون ارزیابی می‌شود.

■ مدل‌سازی وابسته

این وظیفه در واقع تعمیم وظیفه طبقه‌بندی است. در طبقه‌بندی، پیش از آنکه بخواهیم ارزش چندین صفت را برای هدف پیش‌بینی کنیم، مجدداً به کشف قوانین (اگر-آنگاه) پیش‌بینی، برای دستیابی به دانشی سطح بالاتر می‌پردازیم. البته این شکلی عمومی است و هر صفتی می‌تواند هم در قانون مقدم (قسمت اگر) واقع شود و هم در قانون نتیجه (قسمت آنگاه)؛ اما باید توجه داشت که صفت مور نظر نمی‌تواند همزمان در هر دو قسمت قرار گیرد.

■ خوشه‌بندی

همان‌طور که قبلاً ملاحظه شد، در وظیفه طبقه‌بندی، طبقه آموزشی در حکم ورودی الگوریتم داده‌کاوی در نظر گرفته می‌شود و تعیین‌کننده شکلی از یادگیری نظارتی است؛ اما در وظیفه خوشه‌بندی، الگوریتم داده‌کاوی باید با جدا کردن موارد به دسته‌هایی که هر کدام شکلی از یادگیری نظارت‌نشده هستند، خود به کشف روابط و دانش بپردازد. البته تنها یک‌بار که دسته‌ها مشخص شدند، هر خوشه می‌تواند به عنوان یک طبقه در نظر گرفته شود. بنابراین می‌توان یک الگوریتم طبقه‌بندی را بر اساس داده‌ی خوشه‌بندی شده اجرا نمود. (Freitas, 1999) (Park, Song, 1998)

■ کشف قوانین پیوسته

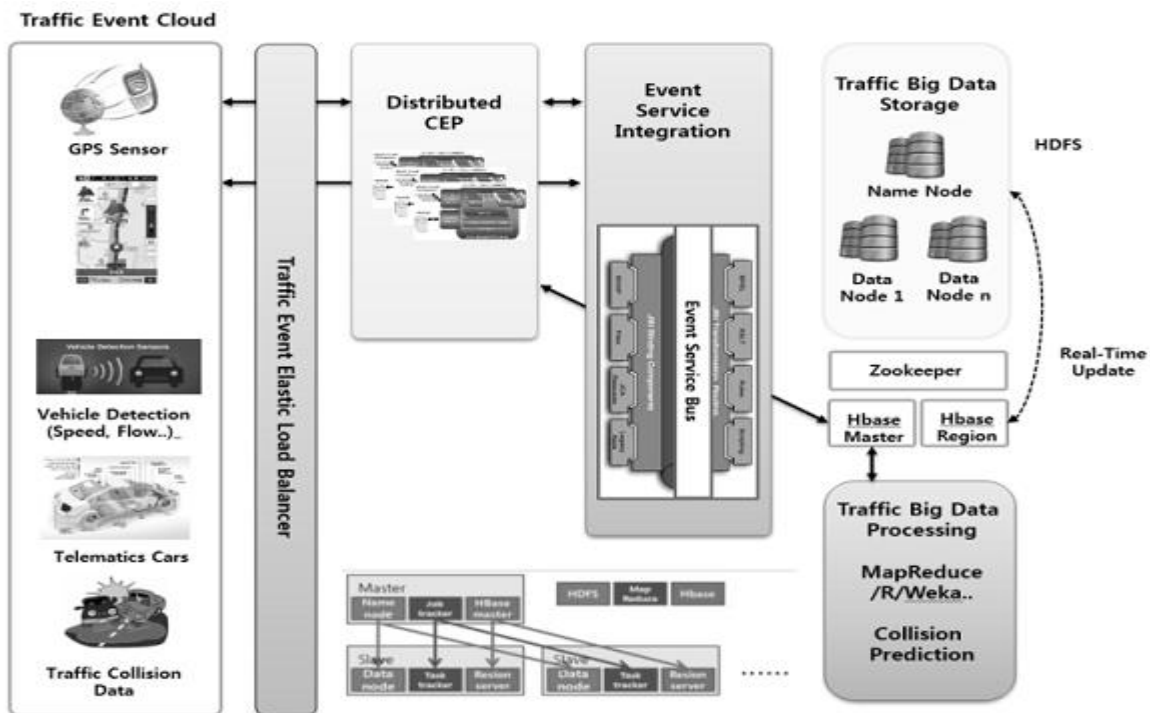
در شکل استاندارد این وظیفه، هر داده‌ی مشاهده‌شده یا ضبط‌شده شامل مجموعه‌ای از صفت‌های دوتایی است که آیتم نام می‌گیرند. گرچه هم طبقه‌بندی و هم قوانین پیوسته، ساختاری (اگر-آنگاه) دارند، از هم متفاوت هستند. باید توجه کرد که در وحله نخست قوانین پیوسته می‌توانند بیشتر از یک آیتم را در قانون نتیجه شامل شوند و این در حالی است که قوانین طبقه‌بندی همیشه دربرگیرنده یک صفت (هدف) هستند. بعلاوه، برعکس وظیفه پیوستگی، وظیفه طبقه‌بندی، قائل به تناسب بین رابطه صفت‌های پیش‌بینی و صفت هدف نیست و صفت‌های پیش‌بینی فقط می‌توانند در قانون مقدم واقع شوند و این در حالی است که صفت هدف صرفاً در قانون نتیجه قرار می‌گیرد (Freitas, 1999).

۹-۲ NoSQL

روند استفاده از پایگاه داده های NoSQL در جایگاه پایگاه داده های ارتباطی در موارد خاص استفاده افزایش یافته است. اغلب الزامات مدل داده می تواند اغلب تغییر کند. در سال ۲۰۱۳، Silva Bastião A Luís و همکاران توضیح می دهند که پایگاه های داده مبتنی بر اسناد محدودیت پایگاه های RDBMS را ندارند [۲۳]. نشان می دهد که Lucene ، MongoDB و CouchDB دارای سطوح عملکرد بالایی هستند.

۱۰-۲ Map Reduction

Map Reduction یک تکنیک است که نقش مهمی در حجم و سرعت داده های بزرگ بازی می کند. مقیاس پذیری الاستیک، افزایش کارایی و در دسترس بودن نقشه نتایج آن است [۱۷، ۱۸].



شکل ۲-۲: معماری پیشنهاد شده برای راه حل بزرگ داده ها [۱]

فصل سوم

پیشینه تحقیق

مقدمه

تحقیقات زیادی در مورد الگوهای ترافیکی در شبکه های جاده ای که محدود به یک شهر و تعداد کمی از جاده ها و / یا اندازه محدودی که در بردارنده سری زمانی هست صورت گرفته است (برخی از مراجع). در این کار، چگونگی تشخیص بسیاری از الگوریتم های مختلف که دارای یک هدف هستند بررسی شده است. شهر دوبلین فرصتی ارائه می دهد که نشان دهیم جاده ها متضاد و دارای تنوع هستند. این بخش با بررسی دقیق پیش بینی و تجزیه و تحلیل ترافیک آغاز می شود. تمرکز بررسی این است روش ها و تکنیک های مختلفی را که محققان در الگوریتم ها، تجزیه و تحلیل با رسانه های اجتماعی و داده های بزرگ استفاده کرده اند، شناسایی کنند.

۳-۱ کارهای پیشین و ادبیات تحقیق

امروز ما در دنیای داده زندگی می کنیم. تحولات در تولید، جمع آوری و ذخیره سازی داده ها، سازمان ها را قادر به جمع آوری مجموعه های داده ای از اندازه های عظیم کرده اند. داده کاوی یک اصطلاح است که روش های تجزیه و تحلیل داده های سنتی با الگوریتم های کشف شده را برای رسیدگی به وظایف این اشکال جدید مجموعه داده ها ترکیب می کند. مقاله [۳۱] یک تحلیل تطبیقی داده های مختلف داده کاوی با استفاده از داده های بزرگ، تجسم و تکنیک های داده کاوی برای پیش بینی و تجزیه و تحلیل ترافیک است. شبکه های حسگر بی سیم یک تکنولوژی است که نقش مهمی را ایفا می کند که شهر های هوشمند شهر با استفاده از این تکنولوژی برای جمع آوری اطلاعات مربوط به ترافیک استفاده می کنند. هدف این است که یک زیرساخت کامل داشته باشیم که نظارت بر رفتارهای ترافیکی را امکان پذیر سازد تا تصمیمات در مورد توسعه شهرها بتواند به شیوه ای دقیق گرفته شود. [۳۱] در حال بررسی کاربرد ابزارهای داده کاوی برای پشتیبانی از پیشرفت دستگاه های قضاوت ترافیک است. رویکرد تجزیه و تحلیل خوشه ای قادر به استفاده از یک توصیف وضعیت سیستم بالا است که از مجموعه وسیعی از سنسورها در یک سیستم سیگنال ترافیکی استفاده می کند

سیستم های اطلاعات مسافرتی پیشرفته^{۱۸} یکی از زمینه های کاربردی سیستم های حمل و نقل هوشمند^{۱۹} است و هدف آن ارائه اطلاعات مسافرتی در زمان واقعی برای مسافران برای تصمیم گیری بهتر مسافرت است. این اطلاعات در صورت ارائه به مسافران در هنگام سفر یا قبل از شروع سفر، موثرتر خواهد بود. بنابراین، مدل های پیش بینی دقیق در ATIS برای انتقال اطلاعات قابل اطمینان در مورد وضعیت آینده ترافیک مورد نیاز است.

روش های مختلفی برای پیش بینی پارامترهای ترافیک استفاده می شود: میانگین میانگین تاریخی، تحلیل رگرسیون، فیلتر کردن کالمن، تحلیل سری زمانی، یادگیری ماشین و غیره. هدف از تحقیق [۲۵]، بررسی استفاده از داده های سنسور خودکار و تکنیک های داده ای برای پیش بینی وضعیت ترافیک تحت شرایط ترافیکی کند زمان سفر و تراکم ترافیک (به عنوان شاخص بارگیری) به طور معمول برای اطلاع رسانی به کاربران در مورد وضعیت یک سیستم ترافیکی استفاده می شود. با این حال، این دو پارامتر به صورت فضایی هستند و اندازه گیری مستقیم از میدان دشوار است.

بنابراین برآورد این پارامترها از داده های مبتنی بر مکان چالشی در بسیاری از پیاده سازی های ITS است. در مقاله [۲۵]، مسئله برآورد تراکم ترافیک با کمک سنسورهای مبتنی بر مکان، قادر به اندازه گیری پارامترهای مانند حجم و سرعت متوسط زمان^{۲۰} است. تکنیک های یادگیری ماشین یعنی k-نزدیک ترین همسایگی (k-NN) و شبکه عصبی مصنوعی (ANN) به عنوان ابزار پیش بینی و ارزیابی در این مطالعه بر اساس عملکرد قابل قبول همان در مطالعات قبلی انتخاب شده اند.

سیستم حسگر از راه دور رایج ترین روش برای نظارت بر ترافیک است. داده های حاصل به شکل حجم ترافیک به جای سرعت یا زمان سفر است. زمان سفر برآورد پایینی از روش معمول. فیلتر کالمن^{۲۱} است، که از پیشرفته ترین روش های در نظریه کنترل مدرن است. این روش ابتدا در سال ۱۹۶۰ توسط Kalman (RE Stephanedes) پیشنهاد شده است. دو روش بسیار خوبی برای پیش بینی جریان و حجم ترافیکی برگرفته شده از نظریه فیلتر کالمن و دیگری UTCS-2 (سیستم کنترل ترافیک شهری) را مقایسه می کند

¹⁸ ATIS

¹⁹ ITS

²⁰ TMS

²¹ Kalman Filtering

[۶]. این مقاله کاربرد های ریاضی را که امروزه در محاسبه سرعت و زمان سفر در کنترل ترافیک شهری استفاده می شود، توضیح می دهد. تحول تکنیک ها برای ارائه خواننده به برخی از روش ها بر روی روش های پیش بینی و سپس با تجزیه و تحلیل دقیق از ۲-UTCS با استفاده از خطای پیش بینی میانگین و خطای متوسط ارائه شده است. در نتیجه بسیاری از شبکه های حسگر بی سیم که در شهرهای مختلف نصب می شوند اندازه گیری می شوند. الگوریتمهایی بر اساس نظریه کالمن برای اندازه گیری حجم زیادی از کنترل حالت های محاسبه زمان سفر ارائه شده اند. این محاسبات ۱۰۰٪ دقیق نیستند، اما یک تکنیک بسیار رایجی هستند که الگوریتم ها طی یک دوره طولانی اصلاح شده و بهبود یافته اند و به عنوان بهترین روش اندازه گیری زمان سفر قابل قبول هستند. دلیل آن حجم سنجی و زمان سفر نیست، بلکه این است که برای سیگنال های ترافیکی و وسایل نقلیه که مسیرها را تکمیل نمی کنند، حساب شود.

۳-۱-۱ سری زمانی پیش بینی

میانگین متحرک خود پیشبینی (ARIMA^{۲۲}) رایج ترین روش برای پیش بینی زمان سفر است. در سال ۱۹۸۳ [۲] تغییرات ARIMA را بیان می کند. (جدول ۲-۲)

تحقیقات در پیش بینی ترافیک یک مورد استفاده معمول در اطراف یک مساله سری زمانی است. الگوریتم هایی که به نظر می رسد بهترین عملکرد را در این منطقه داشته باشند، (ARIMA) و شبکه های عصبی می باشند. به عنوان مثال در سال ۲۰۰۸، Dehuai زنگ و همکاران ال به بررسی تغییرات از مدل خطی ARIMA و غیر خطی کار شبکه، عصبی [۷] و در سال ۲۰۱۰ ادعای مدل رگرسیون بردار (SVR^{۲۳}) حمایت شده است به طور گسترده ای برای حل غیر خطی مشکلات سری های زمانی مورد استفاده قرار گرفت [۸].

این مدل ها به منظور تصادفی بودن عوامل نامشخص به اثبات رسیده اند. این نیز به عنوان ARIMA-GARCH شناخته شده است. GARCH الگوریتم ها و مدل هایی هستند که خطاها را محاسبه می کنند. برخی از عوامل تصادفی مانند حوادث هوایی و جاده ای مورد بررسی قرار گرفته اند [۹-۱۱].

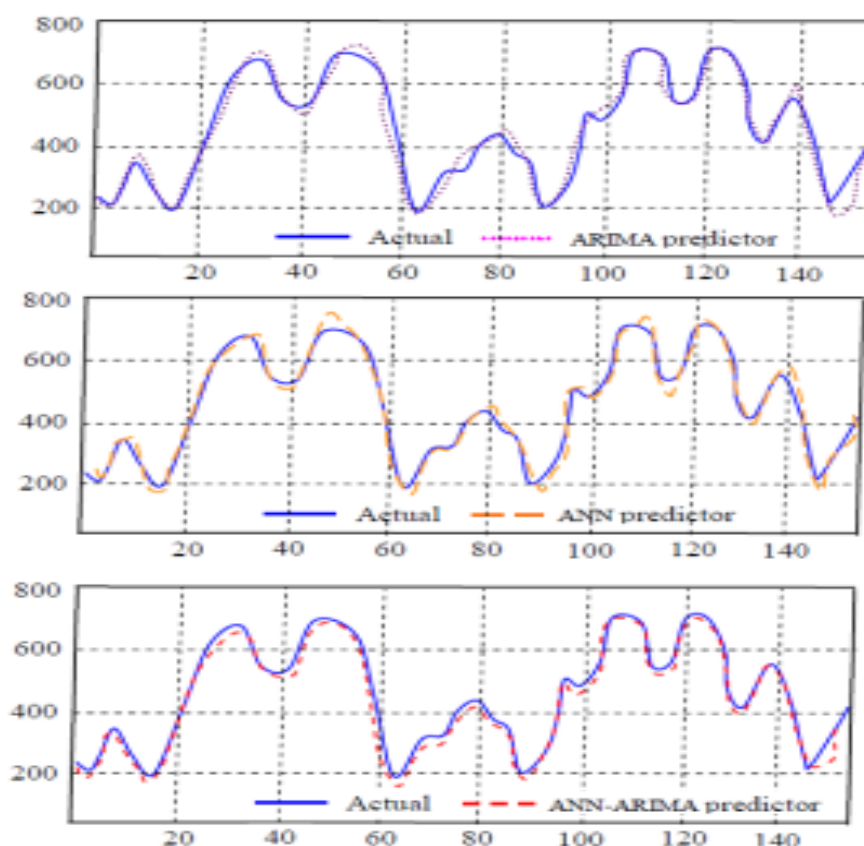
^{۲۲} Autoregressive Integrated Moving Average

^{۲۳} support vector regression

هسته بیشتر تحلیل های پیش بینی ترافیکی با مدل داده های سری زمانی ساخته شده از داده های تاریخی است که توسط Stephanedes (1983) [۱۲] مورد بحث قرار گرفت .

یک مطالعه در سال ۲۰۰۸ Dehuai Zeng et el شبکه عصبی مصنوعی، ARIMA و مدل ترکیبی ANN-ARIMA را مقایسه می کند. (نتایج مطابق شکل ۱-۲ و جدول ۱-۲)

مدل DEHAAI Zeng و همکاران پارامترهایی با مدل BPNN (ARMA) و مدل ترکیبی، بسط یافته BPNN با استفاده از پیش بینی های شرایط خطا برای مدل ARIMA است [۷].



شکل ۱-۳: مقایسه جریان پیش بینی شده سفر با مدل های مختلف

Predictor	rmert%	marent%	rmstert%
ARIMA	0.92	4.26	12.44
BPNN	0.89	3.94	11.64
Hybrid	0.58	2.34	5.68

جدول ۳-۱: مقایسه عملکرد پیش بینی کننده های مختلف

با پیشرفت تکنولوژی، شبکه های جاده ای بهبود یافته و ایمنی خودرو بهبود یافته است. تعداد حوادث جاده ای کاهش یافته و داده های تاریخی را می توان در دسترس قرار داد تا پیش بینی سرعت تاخیر ترافیکی بیشتر از تحقیقات مورد توجه قرار گیرد [۲]. با استفاده از این مدل ARIMA تکامل یافته است. V. Gavirangaswamy et al ARIMA و تغییرات مدل طول می کشد.

SARIMA	Seasonal ARIMA	Good for data with short range recurring pattern
FARIMA	Fractional ARIMA	Considers recurring pattern over long ranges
MARIMA/ARIMAX	Multivariate ARIMA	Includes other time series as dependent variable
k-factor GARIMA	Gegenbauer Polynomials ARIMA	Accounts for both the short-range and long-range dependencies considering different K data frequencies
Switching ARIMA	Different ARIMA models are fitted	Apply different ARIMA for different characteristic

جدول ۳-۲ خلاصه ای از تغییرات ARIMA [۲]

داده های تاریخی مترو دیترویت از ساعتهای بین سال های ۲۰۰۹ تا ۲۰۱۱ جمع شده بودند. نمودار سری زمانی اولیه نشان داد که اطلاعات فصلی که برای SARIMA ایده آل بود. مکانیسم به ثمر رساندن، خطای

متوسط خطی مربعی (RMSE) بود. با استفاده از SARIMA عملکرد آزمون ۵٪ بیش از ARIMA بهبود یافته است. پیش بینی شده مدل ARIMA-GARCH با ۴٪ (جدول ۲-۳) بهبود یافته است. استفاده از این مدل می تواند برای پیش بینی ترافیک کوتاه مدت و آفلاین استفاده شود. مدل محاسبه گرانه در استفاده از برخی از تکنیک ها و تکنیک های داده های جدید برای اجرای این نسل بسیار سودمند خواهد بود [۴].

HV	DS	SH	IH	DS	SH	IH	DS	SH	IH
300	367.29	375.55	143.68	346.52	346.67	140.91	212.82	251.86	89.57
500	374.42	340.71	133.92	361.1	316.53	126.52	214.09	207.26	86.08
800	339.82	355.63	142.42	346.53	347.3	142.33	214.56	207.3	88.33

جدول ۳-۳: مقایسه عملکرد با استفاده از SARIMA، RMSE ARIMA و ARIMA-GARCH با ارزش های تاریخی (HV) در خیابان مرکز شهر (DS)، بزرگراه ایالتی (SH)، بزرگراه بین ایالتی (IH) [۲]

۳-۱-۲ اثرات آب و هوا در ترافیک

در تلاش برای بهبود دقت پیش بینی ترافیک، تحقیقات زیادی برای افزودن متغیرها به داده های ترافیکی تاریخی مانند شرایط آب و هوایی انجام شده است. شکی نیست که شرایط آب و هوایی به نوعی در میزان ترافیک و حجم آن ارتباط دارد. طبق گفته استیفن دوون و بیدیشا قوش در سال ۲۰۱۳ "بارندگی بر شرایط ترافیکی و به نوبه خود حجم ترافیک در شریانی های شهری تاثیر می گذارد" بنابراین در صورت امکان مدل داده ها هنگام ساخت یک الگوریتم پیش بینی برای شرایط ترافیک باید متغیرهای آب و هوایی را شامل شود. استیفن دان و بیدیشا قوش نشان می دهند که با استفاده از نسخه ثابت تبدیل Wavelet گسسته (DWT) به نام SWT برای یک مدل پیش بینی می تواند ارتباط بین حجم ترافیک و شرایط آب و هوایی را که از شبکه های عصبی مصنوعی برای آزمایش های مشابه بالاتر است، نشان دهد. مطالعات حجم داده های ترافیکی را بر اساس KalmanFiltering ایجاد می کنند. یک ساختار SWT برای ایجاد یک سیستم پیش بینی ترافیک نوروواکتیو استفاده می شود. neurowavelet الگوریتم پیش بینی برای جریان ترافیک ساعتی در حالت سطوح بارشی پیشنهاد می کند. این مطالعه از شکل موجک استفاده می کند که در آن تحقیقات دیگر از تغییرات میانگین متحرک استفاده می کند. به طور ایده آل، مقایسه بین میانگین متحرک و

غیره در مقایسه با سایر مدل‌های موجک، ایده آلت‌ر خواهد بود. این مطالعه نشان می‌دهد که بارندگی بر جریان ترافیک تاثیر می‌گذارد [۱۱] .

TCS 106 SWT-ACNN Model		
Overall MAPE	Dry Period	Wet Period
9.0936	10.6463	4.4362
TCS 106 Standard-ANN Model		
Overall MAPE	Dry Period	Wet Period
14.1061	16.5664	6.7254
TCS 125 SWT-ACNN Model		
Overall MAPE	Dry Period	Wet Period
8.0082	9.9116	2.2979
TCS 125 Standard-ANN Model		
Overall MAPE	Dry Period	Wet Period
13.3406	15.9555	5.4958

جدول ۳-۴: مقایسه شرایط مرطوب و خشک در جریان ترافیک

نتیجه فیزیکی یا روند داده‌های ترافیکی را در نظر نمی‌گیرند. حجم ترافیک در روزهای هفته و زمان روز متفاوت است. Keay و Simmonds اثرات متغیرهای آب و هوایی را با حجم جاده در سال ۲۰۰۴ بررسی کردند [۹] . نویسندگان تلاش زیادی در مقایسه داده‌های روند و فصلی در تحلیل نتایج مدل‌های رگرسیون پایه می‌کنند. آنها اطلاعات روزانه و شبانه را برای درک حجم ترافیک و مقایسه آن با داده‌های روزانه تقسیم می‌کنند. آنها همچنین تعداد زیادی از تقسیم‌بندی‌های مختلف را انجام می‌دهند که هر روز دوشنبه تا جمعه و شنبه / یکشنبه جدا می‌شوند و شامل تعطیلات مدرسه و عمومی و غیره هستند. آنها دریافتند Rainfall تاثیر بزرگی در حجم ترافیک دارد. بارش باران بالا و هوای سردتر ترافیک را در شبانه روز کاهش می‌دهد و در ماه‌های سردتر برجسته‌تر می‌شود روزانه با شرایط آب و هوایی مشابه است. دلیل این امر این است که مردم باید به کار و مدارس سفر کنند که فعالیت‌های اختیاری با شرایط سخت تر آب و هوا کاهش می‌یابد. این مطالعه نشان می‌دهد که بین آب و هوای سرد و مرطوب و حجم ترافیک ارتباط وجود دارد. حجم ترافیک در شرایط سرد خنک کاهش می‌یابد. در روزهای هفته کاهش میزان ترافیک در مقایسه با کاهش ۱۷ درصدی در روزهای یکشنبه، حداقل ۱٪ است، که نشان می‌دهد ضرورت برای افراد برای کار و یا مشابه فعالیت‌های عالی است تجزیه و تحلیل‌ها با استفاده از رگرسیون

خطی استاندارد گام به گام در برابر داده های حجم ترافیک فصل، هفتگی و متغیرهای هواشناسی انجام شد [۹].

۳-۱-۳ تکنیک های فضایی

در یک مطالعه ژانگ (۲۰۱۲) از یک روش خوشه بندی ترافیک استفاده می کند تا نقاط جاده ای را که به صورت فضایی و زمان مرتبط هستند، دسته بندی کند. این روش کاهش میزان محاسبه ضروری است. شبکه عصبی مکانیسم پیش بینی پیشنهادی بود. آنها می گویند تحقیقات بعدی برای بهبود دقت لازم است اما تمرکز اصلی این تمرین برای ارائه روش خوشه ای است. آنها الگوریتم خوشه بندی ترافیک آنلاین خود را با خوشه بندی نقطه ترکیبی دینامیکی مشابه پیشنهاد می دهند. این عمل قطعا با در نظر گرفتن یک گزینه اجتناب از هزینه محاسبات بالا در یک راه حل داده پیشنهاد خوبی است. الگوریتم خوشه بندی در مقایسه با شبکه عصبی Bayesian [۱۳] مقایسه شده است .

جاده در شبکه ها هم فضایی و هم زمان دارند. حجم جاده ها بر زمان سفر همسایگان خود تأثیر می گذارد . در جریان بالاتری اهمیت آشکاری دارد و جاده های دور دست ناچیز است. تعدادی از مدل ها برای بهبود ارزش پیش بینی شده در حجم ترافیک آزمایش شده اند . در بسیاری از موارد از مدل پیش بینی سنتی استفاده شده است از جمله Holt Winters و سری زمانی چند بعدی ساختاری. در سال ۲۰۱۲ Yousef- Awwad داراقي etel مقایسه رگرسیون ناووم بیس در برابر مدل های پیش بینی شده. روش پیشنهادی استفاده از یک سری از این وقفه ها تست شده بیش از یک تعداد فواصل زمانی مختلف را با استفاده از گام به گام رو به جلو در حذف و اضافه تعداد متغیرهای می شود را به مدل تا زمانی که تفاوت های بی ربط [۱۴] گنجانده شده است.

۳-۱-۴ رسانه های اجتماعی

Endarnoto و همکاران در سال ۲۰۱۱ مقاله ای با عنوان "استخراج اطلاعات از وضعیت اطلاعات ترافیک و تجسم از رسانه اجتماعی تویتر برای استفاده از (2011) "Android Mobile Application" نوشتند. این تحقیق یک مدل با استفاده از تکنیک های داده کاوی داده ها برای استخراج رویدادهای ترافیکی در جااکارتا ایجاد کرد . "TMC Polda Metro Jaya"، تویتر مرکز ملی مدیریت ترافیک اندونزی. حساب تویتر که

نشان داد عدد داده متناسب با یک متن نیمه ساختار یافته است. در این مورد بخشی از تگ گذاری گفتار نقش مهمی در نتایج دارد. علت اصلی اختلال در نتیجه به علت پیش بینی موقعیت یک مکان "از" یا "به" مکان است. با این حال، آزمایش از یک مدل ساده استفاده کرد که می تواند برای رویداد نه تنها ترافیکی که در آن تاریخ / زمان، مکان به / از و شرایط را حذف می شود استفاده شود. منبع داده های صدای جیر جیر وابسته به کیفیت اطلاعات از کاربر است. در این مورد، گزارشات مربوط به مؤسسه ملی است اطلاعات شهری. این برای بسیاری از شهرها شناخته شده است و به نوبه خود مطالعات مربوطه را در بسیاری از موارد انجام می دهد. این مطالعه از ترتیب متوالی بخشی از نام گفتار استفاده کرد. شکل ۲-۲، براساس فرهنگ لغت POS در جدول ۲-۵ را ببینید.

AT NP V NP AJ V
AT NP V NP N AJ
AT NP V NP N AJ V

شکل ۲-۳: بخشی از الگوی حکم سخنرانی

POS	POS	Name	Example
1	AJ	Adjective	Ramai (crowded), Macet (jammed)
2	AT	Adjective	Time 06:50
3	AV	Adverb	Sangat (highly)
4	CJ	Conjunction	Dan (and), Lalu (then)
5	N	Noun	Lalin (traffic), Arus (stream)
6	NP	Noun	Place Pondok Indah, Bintaro
7	P	Preposition	Di (at), Ke (to), Dari (from)
8	V	Verb	Merayap (crawling), Terjadi (happening)

جدول ۲-۵: بخشی از نام گفتار

مانع اصلی در این تحقیق این است که توییت هایی که با این قوانین مطابقت ندارند، خارج از قوانین و خارج از واژگان هستند و با استفاده از شاخص «POS» انجام می شود. نتایج حاصل از قوانین ساده در بهترین حالت ۷۰٪ از آزمایش شده در کل است. [۱۵]

به جای سر کریسونا Endarnoto و همکاران روش، بی پان و همکاران ال به استفاده از روش TFIDF در برابر توییت ترافیک طبقه بندی شده است. توییت های خود لزوما ترافیکی نیستند بلکه رویدادهای اجتماعی هستند که ممکن است بر روی ترافیک تأثیر بگذارد. در این مطالعه یک چنین رویدادی مشخص شد نمایشگاه رویداد عروسی بود.

این تحقیق همچنین بر مبنای یک منبع از پیش تعیین شده دفتر حمل و نقل پکن برای استخراج ویژگی های مربوط به ترافیک ایجاد می کند. این تحقیق جزئیات محدودی در مورد اینکه چگونه توییت در رویداد مورد استخراج شده یا در صورت استفاده از هر سندی که در آن استفاده می شود، ارائه می شود. توییت از ناحیه ای بود که ناهنجاری های ترافیکی توسط مقامات محلی گزارش شده بود ولی حوادث ترافیکی رخ نداده بودند. [۱۰]

۳-۲ کارهای داخلی

از سال ۲۰۰۰، دوران داده های بزرگ ظهور کرده است. از آن به بعد، برنامه ریزان شهری به طور فزاینده ای از تئوری و روش های داده های بزرگ در برنامه های ریزی شده استفاده می کنند. دهه های اخیر نشان دهنده افزایش سریع استفاده از روش های بزرگ داده در حمل و نقل است، و فرصت های جدیدی را برای نوآوری در مدل سازی حمل و نقل ارائه می دهد. مقاله [۳۲]، نظریه ها و روش های داده های بزرگ را در پیش بینی تقاضای ترافیک تحلیل می کند. با توجه به نظریه، مدل ها و الگوریتم های جدید برای انطباق با داده های جدید بزرگ و پاسخ به محدودیت های روش های تجزیه شده سنتی پیشنهاد می شود. در چنین رویکردهایی، سه روش پیش بینی تقاضای ترافیک، مدل توزیع کامل نمونه تقاضا، مدل ادغام ترافیک، مدل پایگاه داده بیان پروتئین ارگانیزم مورد بحث قرار گرفته است. بدون شک، توسعه داده های بزرگ همچنین چالش های جدیدی را برای روش های پیش بینی تقاضای سفر در زمینه جمع آوری داده ها، پردازش داده ها، تجزیه و تحلیل داده ها و استفاده از نتایج ارائه می دهد. به طور خاص، شناسایی نحوه بهبود رویکردهای پیش بینی تقاضای ترافیک در دوران بزرگ داده در جهان سوم، یک چالش برای محققان در این زمینه است.

در [۳۳] نویسندگان ترافیکی شهری را با استفاده از نرم افزار شبیه سازی Corsim برای شهر تهران استفاده کردند. تحلیل ظرفیت و بهینه سازی جریان ترافیک در خیابان دستغیب حدفاصل استاد معین و بزرگراه

آیت اله سعیدی به وسیله نرم افزار شبیه سازی Corsim، مورد بررسی قرار گرفت. در شبیه سازی این شبکه، اطلاعات فیزیکی و ترافیکی مورد نیاز مربوط به داده های ورودی از اطلاعات فایل GIS سازمان حمل و نقل و ترافیک و برداشت میدانی حاصل شد.

حجم تصادفات ترافیکی منطقه ۸ تبریز در شبکه پیچیده شهری به روش برآورد چگالی هسته (KDE) با استفاده از ARCMAP و نرم افزار SANET در [۳۴] مورد تجزیه و تحلیل قرار گرفته است.

برای تحلیل داده های ترافیکی خودروها در تمامی مسیرهای شهری از خودروهای مجهز به حسگرهای سیستم موقعیت یاب جهانی (GPS) که در مسیرهای حمل و نقل شهری توزیع و اطلاعات مکانیشان ارسال می گردد، میتوانند ابزار خوبی برای تولید داده های ترافیکی در سطح شهر باشند. وسایل نقلیه ی مجهز به حسگر بدلیل میزان کارایی، بهره وری و دقت بالا منبع مهمی از اطلاعات حرکت خودر وها در سیستم های حمل و نقل هوشمند (ITS 2) هستند که به موجب آن، در مقاله [۳۵] از داده های سنسور موقعیت یاب نصب شده روی تاکسی ها که در فواصل زمانی معینی داده ارسال می نماید، استفاده شده است .

۳-۳ داده های بزرگ

امروز ما در دنیای داده زندگی می کنیم. تحولات در تولید، جمع آوری و ذخیره سازی داده ها، سازمان ها را قادر به جمع آوری مجموعه های داده ای از اندازه های عظیم کرده اند. داده کاوی یک اصطلاح است که روش های تجزیه و تحلیل داده های سنتی با الگوریتم های کشت شده را برای رسیدگی به وظایف این اشکال جدید مجموعه داده ها ترکیب می کند. مقاله [۲۶] یک تحلیل تطبیقی داده های مختلف داده کاوی با استفاده از داده های بزرگ، تجسم و تکنیک های داده کاوی برای پیش بینی و تجزیه و تحلیل ترافیک است. شبکه های حسگر بی سیم یک تکنولوژی است که نقش مهمی را ایفا می کند که شهر های هوشمند شهر با استفاده از این تکنولوژی برای جمع آوری اطلاعات مربوط به ترافیک استفاده می کنند. هدف این است که یک زیرساخت کامل داشته باشیم که نظارت بر رفتارهای ترافیکی را امکان پذیر سازد تا تصمیمات در مورد توسعه شهرها بتواند به شیوه ای دقیق ساخته شود. [۲۶] در حال بررسی کاربرد ابزارهای داده کاوی برای پشتیبانی از پیشرفت دستگاه های قضاوت ترافیک است. رویکرد تجزیه و تحلیل خوشه ای قادر به

استفاده از یک توصیف وضعیت سیستم بالا است که از مجموعه وسیعی از سنسورها در یک سیستم سیگنال ترافیکی استفاده می کند

داده های بزرگ اطلاعاتی هستند که بزرگی و پیچیدگی در آن بسیار مشکل ساز است. این موارد عمدتاً بر V۴ در اطلاعات بزرگ تمرکز دارد [۱۶] .

- حجم

در سال ۲۰۰۰ یک کامپیوتر ممکن بود ۱۰ گیگابایت حافظه داشته باشد. سایت های رسانه های اجتماعی مانند توییتر و فیس بوک روزانه ۵۰۰ ترابایت مصرف می کنند.

- سرعت

این بیشتر مربوط به گرفتن اطلاعات زمان واقعی در سرعت بالا است. به طور خاص توییتر یک مثال خوب از نظارت بر داده ها در زمان واقعی است. همچنین به عنوان مصرف پیامهای واقعی از طرف کاربران، آنها API ها را افشا می کنند که اجازه می دهد تا اهرم های عمومی در این مورد استفاده شود . مصرف داخلی توییتر به سرعت پردازش همان اندازه و یا حتی بیشتر از ۵۰۰ terrabytes از داده ها است.

- تنوع

داده های بزرگ باید بتوانند انواع داده های مختلف مانند ویژگی های فضایی، گرافیکی، صوتی و تصویری و متن بدون ساختار را اداره کنند. Traditional RDBMS برای اندازه گیری حجم کمتر داده های ساختاری طراحی شده است.

- اطمینان

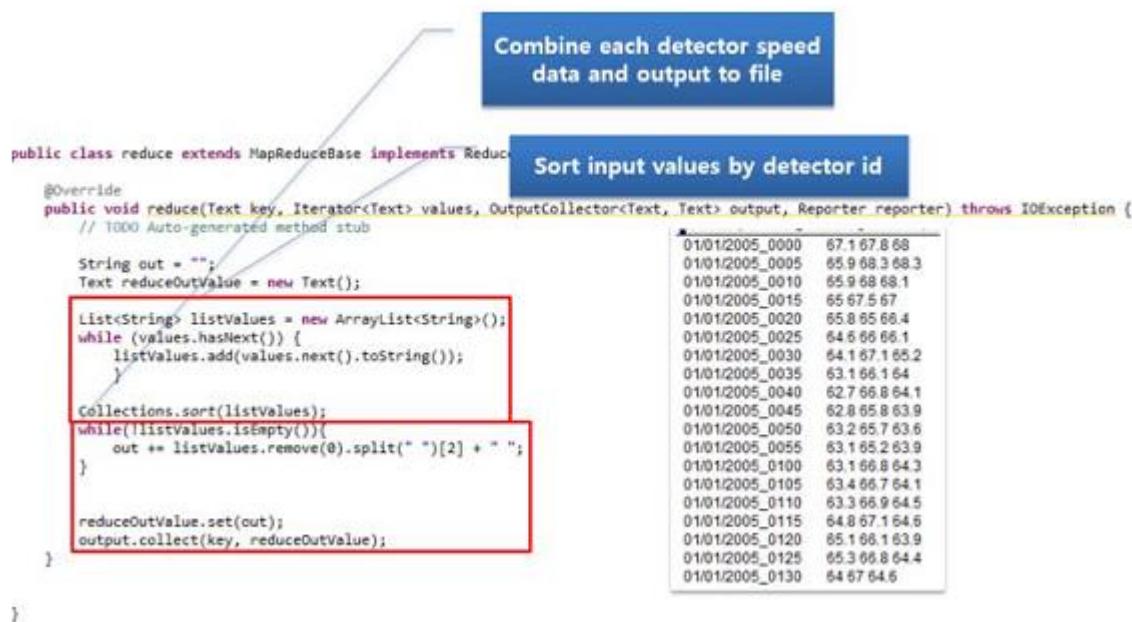
اظهارنظر جدیدی از V است. با اطمینان از کیفیت داده ها آیا اصطلاح استفاده از تجزیه و تحلیل داده ها برای تصمیم گیری، حل مسئله و نتایج دانش موثر است..

سیستم های پایگاه داده سنتی برای کار بر روی یک ماشین طراحی شده اند. که محدودیتی برای مقیاس پذیری راه حل به عنوان ظرفیت محدود است. استفاده از شیوه های کاربردی و توسعه یافته، به عنوان

فضای ابری برای پایگاه های کاربردی چند پایه تکامل یافته اند بطوریکه نیاز به رشد پایگاه داده برای بیشتر کاربران با استفاده از سیستم وجود دارد. پایگاه داده های بزرگ داده، مانند MongoDB، این مشکلات را حل می کنند و به شرکت ها کمک می کنند تا ارزش کسب و کار عظیم ایجاد کنند [۵].

۳-۴ کاهش نقشه

در بعضی از آثار ذکر شده در این کار، یک مشکل رایج، جزئیات با حجم زیاد داده ها از اطلاعات ترافیکی و اطلاعات توییت [۲،۷،۱۰] است. در سال های اخیر اصطلاح "بزرگ داده ها" مفید بوده است. VinayGavirangaswamy و همکاران [۲] با توجه به آزمایشی در حدود ۲۲۰ ساعت محاسباتی انجام دادند تا این آزمایشات بر روی یک ماشین با ۸ گیگابایت RAM انجام شود. داده های بزرگ اطلاعاتی است که بزرگی و پیچیدگی در آن مشکل ساز می شود. بریتو و همکاران رویکردی به نام جریان MapReduce الکترونیکی را به عنوان یک راه حل برای مشکل بزرگ داده پیشنهاد دادند.



شکل ۳-۴ : نقشه و کاهش [۱]

From	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	1.98E-05	0	0	0	0	1.73E-05	0	0	0	0	0	0	0	0	0
10	0	2.24E-05	5.60E-06	6.31E-06	0	2.90E-05	3.44E-05	2.09E-05	0	0	1.88E-05	2.77E-05	0	0	0	0
15	0	4.91E-05	0	7.12E-06	5.71E-06	8.48E-06	3.31E-06	7.28E-06	3.70E-06	0	1.44E-05	0	0	2.07E-05	0	0
20	0	0	5.71E-06	4.61E-06	5.03E-06	4.82E-06	5.64E-06	3.33E-06	3.07E-06	6.89E-06	0	1.01E-06	0	1.01E-06	0	0
25	0	0	0	0	3.41E-06	3.91E-06	4.52E-06	2.81E-06	6.32E-06	2.45E-06	8.45E-06	1.91E-06	7.84E-07	5.32E-06	2.34E-06	0
30	0	0	0	1.95E-06	5.22E-06	3.87E-06	4.63E-06	3.75E-06	3.78E-06	3.46E-06	4.13E-06	2.31E-06	3.50E-06	2.78E-06	2.26E-06	0
35	0	0	0	2.09E-06	6.33E-06	4.73E-06	3.74E-06	2.33E-06	2.62E-06	1.37E-06	3.36E-06	1.00E-06	2.88E-06	2.18E-06	2.80E-06	0
40	0	0	0	8.78E-06	5.93E-06	5.88E-06	1.04E-06	4.38E-06	5.80E-06	4.05E-06	2.24E-06	2.38E-06	2.18E-06	1.07E-06	2.05E-06	0
45	0	0	1.46E-06	6.70E-06	4.97E-06	4.46E-06	4.91E-06	3.72E-06	2.28E-06	2.01E-06	1.58E-06	1.55E-06	1.88E-06	4.12E-06	0	0
50	0	0	0	2.71E-06	5.85E-06	4.94E-06	2.12E-06	3.43E-06	3.38E-06	1.80E-06	1.84E-07	1.14E-07	0	1.05E-06	0	0
55	0	0	0	1.23E-05	7.53E-06	3.02E-06	1.29E-06	2.65E-06	3.83E-06	2.03E-06	3.87E-07	2.24E-07	1.11E-07	1.18E-07	4.19E-07	0
60	0	0	2.36E-05	0	4.58E-06	1.13E-06	4.36E-06	3.93E-06	7.38E-07	1.41E-06	7.69E-07	4.32E-07	1.24E-07	2.88E-07	1.91E-07	0
65	0	0	0	1.81E-05	3.84E-06	1.95E-06	3.31E-06	1.69E-06	1.79E-06	1.36E-06	1.88E-07	2.57E-07	1.84E-07	1.18E-07	1.84E-07	0
70	0	0	0	0	0	2.11E-05	5.27E-06	0	0	0	0	4.48E-07	1.28E-07	1.08E-07	2.18E-07	0
75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

شکل ۳-۵: پیش بینی برخورد [۱]

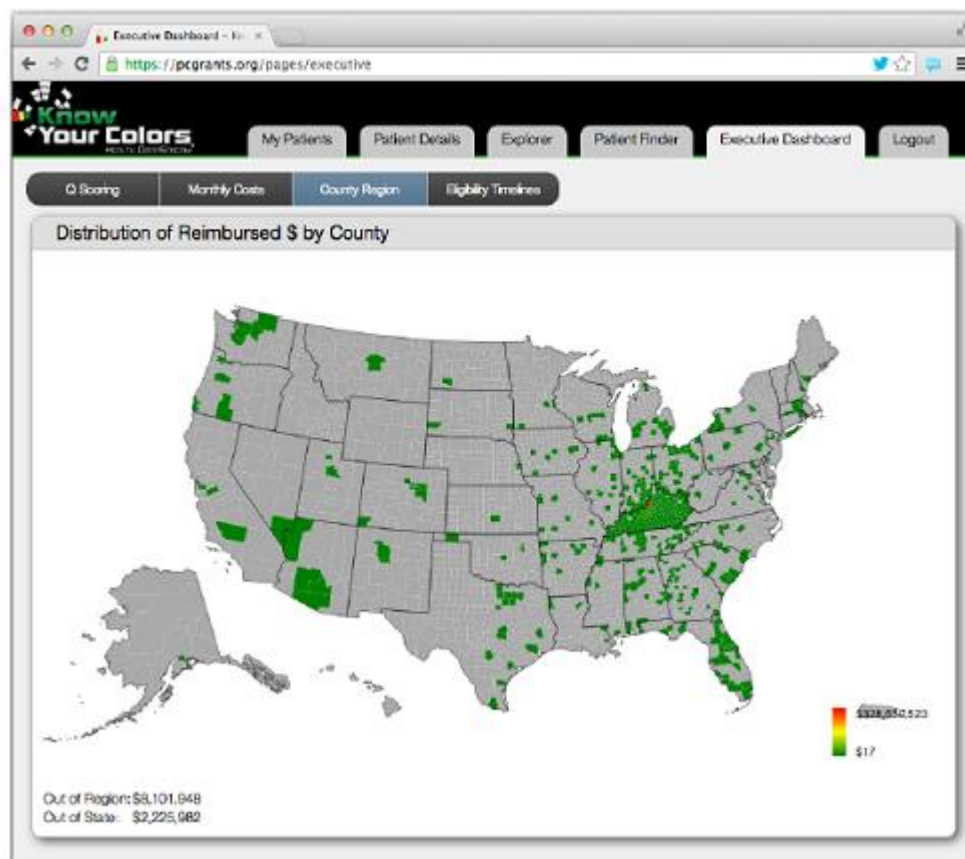
ویژگی های داده های بزرگ عمدتاً حجم، تنوع، سرعت و صحت است [۱۹، ۲۰]. تحقیقات ادعا می کند که مکانیزم آن می تواند صد برابر در زمان پاسخ دهی بهبود یابد و افزایش ده برابر در هر گره در مقایسه با Hadoop را ارائه می دهد. در اغلب موارد، بهبود در عملکرد با پایین آمدن حجم داده ها است که به ناچار برخی از داده ها و یا نقشه برداری اسناد که حاوی داده های همان کلاس را از دست بدهند. این مطالعه نشان می دهد که MapReduce و / یا پردازش جریان رویداد به تمام مشکلات داده بزرگ جواب نمی دهد و جریان MapReduce یک راه حل برای برخی از موارد است [۲۱]. با این حال در سال ۲۰۱۳ Duckwon Chung et al تکنولوژی داده های بزرگ Hadoop و HBase را برای تجزیه و تحلیل ترافیک در زمان واقعی از تعدادی از منابع مختلف، اطلاعات ترافیکی، سایت های اجتماعی، سیگنال های GPS تلفن همراه مورد استفاده قرار می دهند. راه حل شامل چندین گره داده برای مصرف داده های مشاهدات توزیع شده توسط یک گره داده اصلی است (شکل ۳-۳). گره اصلی تصمیم می گیرد که کدام یک از گره ها برای بررسی مشاهدات بر روی یک شاخص را ارسال کند. در این مورد موقعیت مکانی توسط آشکارسازها بیشترین شاخص را در نظر می گیرند. وگره های پس از آن نقشه و تاریخ را کاهش

می دهند که یک روش جمع آوری داده ها برای تجزیه و تحلیل های بیشتر است. هنگامی که الگوریتم پردازش تجمع را می توان تولید کرد پیش بینی برخورد (شکل ۳-۴ و ۳-۵) [۱].

به طور کلی قابل درک است که تجزیه و تحلیل جریان حجم داده های مصرف شده توسط یک برنامه توییتتر است. با استفاده از یک داده بزرگ توییتتر McCreadie و همکارانش با تکنیک تقسیم و تسخیر به منظور مقادیر جریان داده های بزرگ در هزاران توییت بر ثانیه آزمایش می کنند [۴]. McCreadie به نظر می رسد که MapReduce و DBMS های سنتی را برای پردازش در حال اجرا در توییتتر در زمان واقعی مناسب می داند، به خصوص DBMS که در آن از روش "ذخیره و پس فرآیند" برای برخورد با داده های بزرگ استفاده می کند. این اپلیکیشن از یک پلتفرم به نام Storm استفاده می کند که در حال حاضر بخشی از پشته Hadoop است که جریان داده های واقعی را در اختیار دارد. در طی آن Topology Detection Event، از الگوریتمی برای خوشه بندی داده ها استفاده می کند که از طریق Split Key Partitioning (DLKP)) برای دسته بندی اسناد داده ها در گروه های مشابه است. DLKP اصطلاحی برای فاصله محلی که کوزینس این مقاله را روشی مناسب برای مدیریت داده های توییتتر بدون ساختار با ویژگی های کلیدی ناشناخته می داند. [۴]

۳-۵ داشبورد تحلیلی

در Big Data همه چیز در مورد نوشتن و خواندن داده ها نیست. لازم است که دیدگاه های تحلیلی داده ها را ارائه دهیم. برای کاربران خواندن داده های حجیم دشوار است. داشبورد تحلیلی تکنیک نمایش اطلاعات تحلیلی است. در سال ۲۰۱۳، کریستوفر رایز و همکاران، اهمیت استفاده از داشبورد بصری برای تحلیل مقادیر زیاد اطلاعات را توضیح می دهند [۲۲]. مثال هایی برای نشان دادن چگونگی استفاده بهتر از رنگ ها و اطلاعات فضایی ارائه شده است، شکل ۳-۶ و ۳-۷ را ببینید.



شکل ۳-۷: تجسم فضایی داده های بزرگ

دیگر الگوریتم های غیر خطی می توانند در مدل های رگرسیون به عنوان رگرسیون بردار پشتیبانی استفاده شوند تا زمانی که کرنال به RBF برسد و تقریب ماته کارلو از تبدیل فوریه آن را تنظیم کند، در حالی که می تواند تبدیل گرادیانی تصادفی (SGD) هنگام استفاده از آن به عنوان یک شبکه عصبی مصنوعی باشد. همانطور که می دانیم شبکه عصبی که به آرامی انجام می شود، SGD به خوبی برای یادگیری های بزرگ به کار رود [۲۳، ۲۴].

در [۲۷] نویسندگان ترافیکی شهری را با استفاده از نرم افزار شبیه سازی Corsim برای شهر تهران استفاده کردند. تحلیل ظرفیت و بهینه سازی جریان ترافیک در خیابان دستغیب حدفاصل استاد معین و بزرگراه آیت اله سعیدی به وسیله نرم افزار شبیه سازی Corsim، مورد بررسی قرار گرفت. در شبیه سازی این شبکه،

اطلاعات فیزیکی و ترافیکی مورد نیاز مربوط به داده های ورودی از اطلاعات فایل GIS سازمان حمل و نقل و ترافیک و برداشت میدانی حاصل شد.

حجم تصادفات ترافیکی منطقه ۸ تبریز در شبکه پیچیده شهری به روش برآورد چگالی هسته (KDE) با استفاده از ARCMAP و نرم افزار SANET در [۲۸] مورد تجزیه و تحلیل قرار گرفته است.

برای تحلیل داده های ترافیکی خودروها در تمامی مسیرهای شهری از خودروهای مجهز به حسگرهای سیستم موقعیت یاب جهانی (GPS) که در مسیرهای حمل و نقل شهری توزیع و اطلاعات مکانیشان ارسال می گردد، میتواند ابزار خوبی برای تولید داده های ترافیکی در سطح شهر باشند. وسایل نقلیه ی مجهز به حسگر بدلیل میزان کارایی، بهره وری و دقت بالا منبع مهمی از اطلاعات حرکت خودر وها در سیستم های حمل و نقل هوشمند (۲) ITS هستند که به موجب آن، در مقاله [۲۹] از داده های سنسور موقعیت یاب نصب شده روی تاکسی ها که در فواصل زمانی معینی داده ارسال می نماید، استفاده شده است.

۳-۶ الگوریتم ها

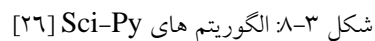
الگوریتم های خطی و غیر خطی برای پیش بینی سری زمانی استفاده شده اند. دکتر Vincent Granville در سال ۲۰۱۴ الگوریتم های رگرسیون خطی را توصیف می کند و در فهرست زیر خلاصه می شود [۲۴].

- رگرسیون خطی
- قدیمی ترین مدل رگرسیون است. حساس به over-fitting و outliers می باشد.
- لجستیک (Poisson یا Cox) رگرسیون
- اغلب در کارآزمایی های بالینی ، به دست آوردن امتیاز و شناسایی تقلب مورد استفاده قرار می گیرد و در نظر گرفته می شود.
- رجیج رگرسیون (Ridge)
- رگرسیون با محدودیت در ضرایب. به عنوان مدل رگرسیون خطی بیش از حد حساس نیست.
- رگرسیون لسو (Lasso)
- همانند Ridge به جز آنکه به طور خودکار از کاهش متغیر استفاده می کند.

●

●

●



۷-۳ نتیجه گیری

کنترل فضایی حالت روش اصلی برای اندازه گیری حجم و سرعت ترافیک است. نظریه کالمن برجسته ترین الگوریتم های امروز است و به طور گسترده ای در سیستم های نظارت بر ترافیک استفاده می شود. این روش ها دقیق مطلق را تضمین نمی کند، بلکه بیشترین استفاده را در سیستم های برآورد ترافیک دارد. بر اساس داده های تولید شده از سیستم های نظارت، روش های پیش بینی برای پیش بینی ترافیک اجرا می شوند. بسیاری از مدل مورد استفاده برای پیش بینی ARIMA و GARCH است. GARCH برای مدیریت سر و صدا در مجموعه داده های ترافیکی استفاده می شود. برای هدف الگوریتم های ورزش از Python SciPy مقایسه خواهد شد. Python SciPy دارای طیف گسترده ای از مدل های خطی است، برخی از آنها سر و صدا را به روش های مختلف و فقط گزینه های محدود غیر الگوریتم خطی. پروپترون برای شبکه عصبی مصنوعی و رگرسیون برداری (SVR) مورد استفاده قرار خواهد گرفت.

فصل چهارم

روش پیشنهادی

مقدمه

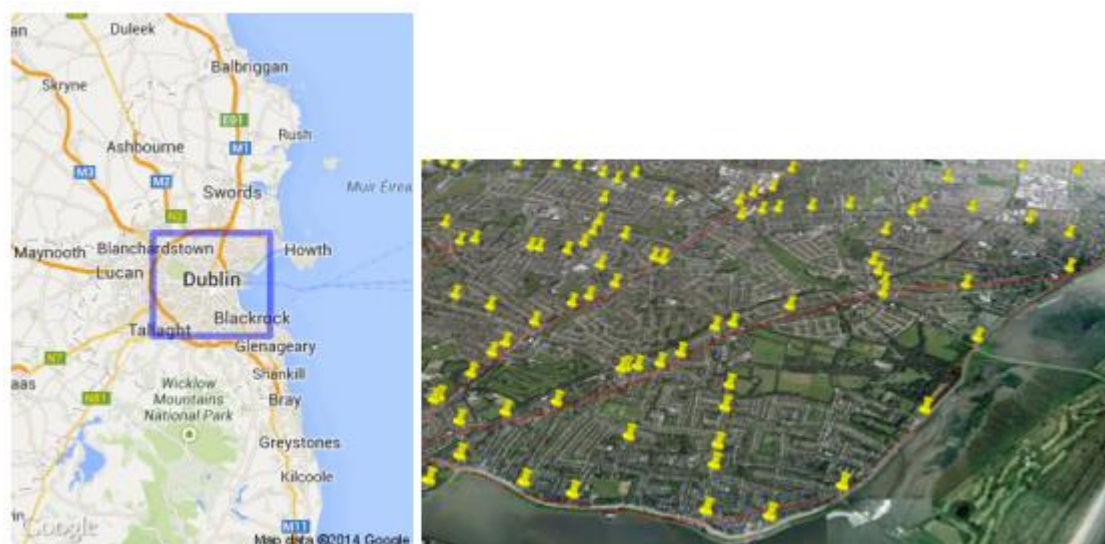
در این بخش ابتدا اطلاعات موجود از منابع باز قبل از جمع آوری داده ها ارائه شده است. هدف از این قسمت ارائه پیش زمینه ای در مورد اطلاعات موجود است. سپس روشی که چگونه از تکنیک های داده بزرگ برای ذخیره داده ها در یک بانک اطلاعاتی بدون ساختار استفاده می شود ذکر می گردد. برای تجزیه و تحلیل ترافیک مورد بررسی قرار می دهیم.

۴-۱ مجموعه داده های ترافیک

داده ها از سه مجموعه داده تشکیل شده است که از طریق وب سایت داده باز^{۲۴} DubLinked در دسترس قرار می گیرد. بارهای سفر برخی از مسیرهای اصلی در سراسر شهر دوبلین ارائه شده است. هر مسیر از تعدادی لینک تشکیل شده است ، هر لینک یک جفت سایت کنترل ترافیک جغرافیایی ارجاع شده است. DubLinked نقشه هایی را توزیع می کند که می توانند به Open Street Map و Google Maps معروف به shapefiles و پرونده های KML وارد شوند ، به شکل زیر مراجعه کنید.

با این کار مکانهای کنترل ترافیک توسط پین های زرد در نقشه های Google در امتداد مسیرهای مشخص شده مشخص می شوند. هدف از سایت های کنترل ترافیک نظارت بر میزان ترافیک با استفاده از سنسورها است.

^{۲۴} <https://data.smartdublin.ie/dataset>



شکل ۴-۱: نقشه گوگل دوبلین

۴-۱-۱ اطلاعات ترافیکی

داده های ترافیکی تاریخی از طریق فهرست بایگانی DubLinked TRIPS ذخیره می شوند [۳]. بایگانی یک لیست دایرکتوری HTTP از فایل های فشرده باینری است. هر پرونده فشرده دودویی حاوی یک روز داده های مشاهده تاریخی با فرمت CSV 3.2.1 است. نام پرونده با تاریخ در قالب روز - YYMMDD.csv.bz2 مشخص شده است.

Attribute	Description
Timestamp	Date time of observation YYYYMMDD-HHmm
Route	Road with 1 or more observed links
Link	Segment of road between 2 control sites
Direction	Direction of flow of traffic

جدول ۴-۱ متغیرهای ثبت شده در داده های ترافیکی

۴-۱-۲ داده اتصال

DubLinked داده های اتصال را که مربوط به سایت های کنترل ترافیک است، ارائه می دهد. هر یک از موارد ثبت شده شناسه های نگهدارنده را در مورد دو سایت کنترل ترافیک به نام TCS1 و TCS2 ثبت می

کند که در لیست زیر مشاهده می شود. جزئیات در سه قالب ، CSV, KML و پرونده Shape ارائه شده است که نمونه آن در 3.2.2 junctions.csv است. بین سایت های کنترل ترافیک TCS1 و TCS2 مقدار زمان سفر تخمین زده شده است و در منابع بیشتر در تحقیق به عنوان مکان مشاهده شده (OL) شناخته می شود.

<i>Attribute</i>	<i>Description</i>
SiteID	Relates to the identifier TCS1 and TCS2
X	Longitude [Irish Grid (IG; EPSG:29902) Coordinate Value]
y	Latitude [Irish Grid (IG; EPSG:29902) Coordinate Value]
Location	Name of Road

جدول ۴-۲ متغیرهای جدول اتصال

۴-۱-۳ داده های مسیرها

DubLinked داده های مسیر را ارائه می دهد که مربوط به بخشی از جاده است که از تعدادی لینک تشکیل شده است. پیوند طول جاده بین سایتهای کنترل ترافیک (TCS1 و TCS2) است و در بالا به آن اشاره شد. هر مشاهده ثبت شده مسیر و لینک شناسه نگه داشته ، جدول زیر را ببینید. جزئیات در سه قالب ، CSV ، KML و Shapefile با مسیرها ارائه شده است. در لیست زیر مشاهده کنید.

<i>Attribute</i>	<i>Description</i>
Route	Stretch of road being monitored
Link	Segment of the Route
Direction	Direction of traffic along link
TCS1	Control point for traffic entering link
TCS2	Control point for traffic exiting link
WKT	Irish Grid Coordinates

جدول ۴-۳ متغیرهای داده های مسیر

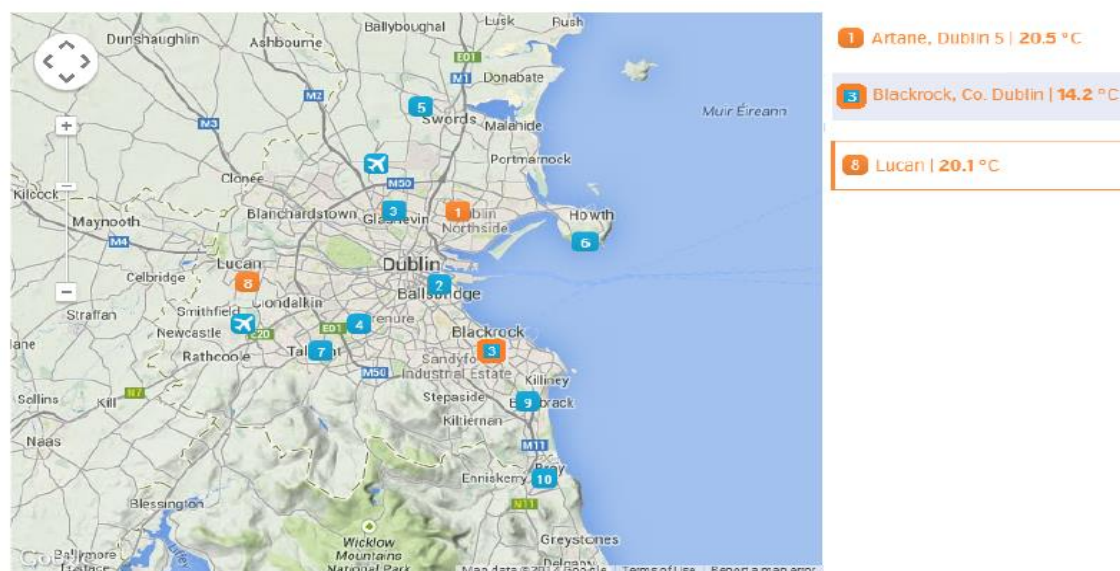
۴-۲ اطلاعات آب و هوا

تحقیقات زیادی نشان داده اند که تأثیر شرایط آب و هوایی بر روی ترافیک چه بوده است. در این بخش فرایند استخراج داده های آب و هوا پوشش داده می شود. داده های هوا به خودی خود از یک وب سایت منبع باز (Wunderground Weather) گرفته شده است.

Wunderground ارائه دهنده داده های ایستگاه هواشناسی است. ایستگاه های هواشناسی متعلق به عموم مردم است. این ایستگاه های دیافراگم به عنوان منبع داده های هواشناسی انتخاب می شوند ، شرایط آب و هوا با گذشت زمان تغییر می کند. شرایط آب و هوای مرطوب می تواند در یک زمان خاص برای یک منطقه کوچک باشد و مساحت بیشتر دوبلین را با شرایط مرطوب تجربه یکباره تضمین نمی کند. به عنوان مثال مسافرت در باران زمان می برد. این بدان معنی است که باران باعث ترافیک در زمان ها و مکان های مختلف می شود. ایستگاه هواشناسی در شمال ، غرب و خارج از شهر دوبلین واقع شده است.

<i>Id</i>	<i>Location</i>
ICODUBLI2	Lucan, Co Dublin West
ILEINSTE8	Blackrock, Dublin 8, South
IDUBLINC2	Artane, Dublin 5, North

جدول ۴-۴ متغیرهای اطلاعات آب و هوایی



شکل ۴-۲: مکانهای ایستگاه هواشناسی

نمادهای شماره در شکل ۳-۲ ایستگاه های هواشناسی موجود در وب سایت Wunderground^{۲۵} دسترسی به آب و هوای تاریخی هر روز و مکان جداگانه از طریق وب را فراهم می کند [۲۷]. نقاط برجسته به رنگ نارنجی تصرف ایستگاه هواشناسی برای اهداف پایان نامه است.

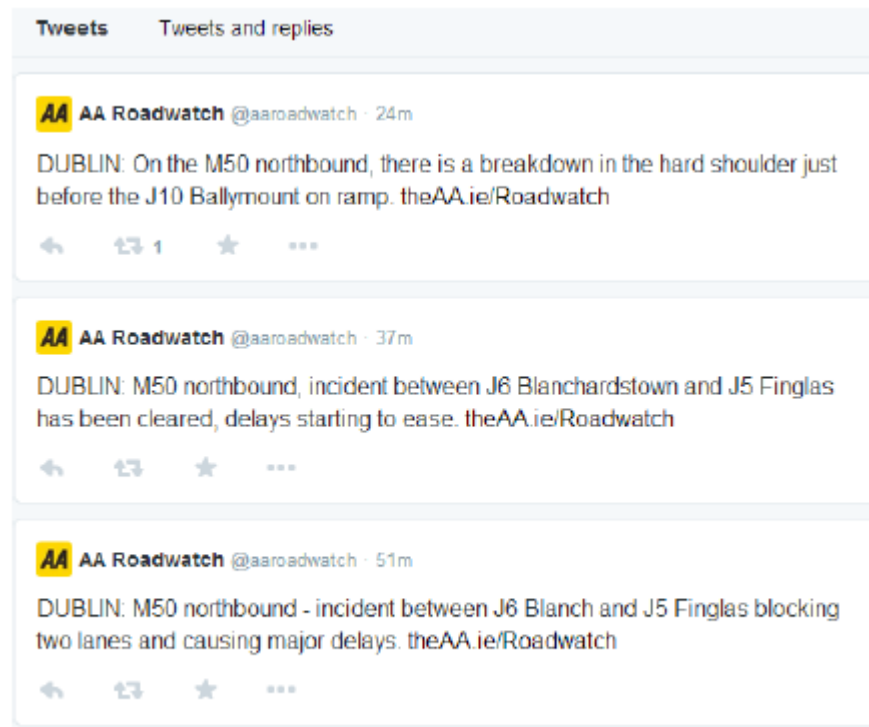
۴-۳ داده های تویتر

تویتر API را برای جستجوی داده های تاریخی بر اساس انواع مختلف فیلترها در تویتهای ویژگی و / یا جستجوی جریان فراهم می کند. با استفاده از تکنیک گرفتن اطلاعات از ارائه دهنده تویتهای مربوط به ترافیک برای طبقه بندی تویتهای پخش مستقیم. ارائه دهندگان شناخته شده داده های ترافیک مانند AA Roadwatch 3.5 یک سرور عمومی برای ارائه به روزرسانی اطلاعات راهنمایی و رانندگی هستند.

۴-۴ داده های ترافیک جدول زمانی کاربر تویتر

تویتهای ارائه دهنده AA Road Watch و Live Drive هیچ اطلاعات جغرافیایی دیگری را با داده های متنی ندارند. جایی که جریان زنده اطلاعات مربوط به اطلاعات جغرافیایی را انجام می دهد یعنی هماهنگ می شود. اما داده های متنی ممکن است مکان را فراهم نکند. هدف در اینجا پیوند دادن تویتهای مربوط به ترافیک با موقعیت جغرافیایی در توییت است. تفاوت این دو منبع در این است که توییت های Live Drive دوباره توییتهای از طرف عموم مردم باز می شود و توییت های AA Road Watch توییتهای است که به عنوان یک سرویس National ارائه می شود. اگرچه توییت های AA Road Watch و Live Drive مربوط به ترافیک هستند ، اما این تضمین نمی کند که همه توییت ها حاوی اطلاعات مربوط به ترافیک هستند [۴].

²⁵ <http://www.wunderground.com/personal-weather-station/dashboard?ID=IDUBLINF2#history/data/s20140423/e20140423/mtoday>

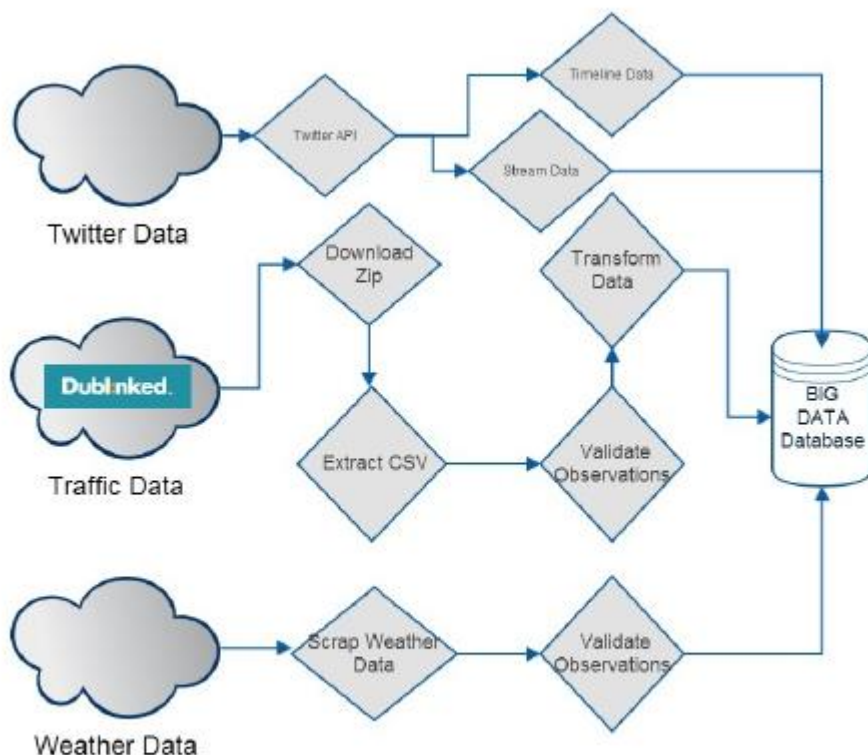


شکل ۴-۳: نمونه ای از توییتها

هدف استفاده از این دو منبع برای توییت این است که از توییت‌های مخصوص کاربر می‌توان برای استخراج ویژگی‌هایی که مربوط به موضوع توییت‌های ترافیکی هستند استفاده کرد و از یک اندازه‌گیری شباهت با یادگیری قانون همراه برای مطابقت با توییت‌های مربوط به ترافیک از زمان واقعی استفاده کرد. داده‌های حاوی موقعیت جغرافیایی که جزئی از داده‌های زمان بندی کاربر نیست.

۴-۵ جمع‌آوری داده‌ها

داده‌های این تحقیق شامل سه حوزه اساسی ترافیک، آب و هوا و داده‌های توییت‌ها است. تمام داده‌ها را می‌توان از طریق وب و داده‌های باز روی خط دریافت کرد به علت عدم دسترسی به داده‌های شهر تهران این پایان‌نامه بومی سازی نشد. در تکنیک‌های مورد استفاده برای جمع‌آوری و ذخیره‌سازی داده‌ها برای این سه بخش بصورت شکل ۴-۳ می‌باشد.



شکل ۴-۴: نحوه جمع آوری داده ها

در نتیجه این مجموعه ، تمام سوابق مشاهدات به دست آمده و بدون از دست دادن داده در دسترس هستند. هر مشاهده مربوط به ترافیک می تواند دوباره به شکل کامل آن بازگردد. همین مورد در مورد توییت های توییت نیز صدق می کند. توییتها به مجموعه دیگری حذف شده و صفات غیر ضروری را برای جستجوی سریع تحلیلی حذف می کنند. هر رکورد را می توان به شکل اصلی خود برگرداند. در صورت نیاز به تحقیقات بیشتر برای یک کاربر توییت ، این ممکن است مفید باشد.

۶-۴ اکتشاف داده ها

در بخش اکتشاف داده ها توضیحات مفصلی از داده های جمع آوری شده در بخش قبل به همراه تجسم ارائه می شود. با استفاده از تجسم از طریق نقشه های گوگل و پایتون ، برای شناسایی موضوعات با کیفیت استفاده می شود که فیلترها از طریق یک فرآیند دستی یا استفاده از ابزارهای استاندارد امکان پذیر نیست.

هدف از این بخش تولید یک مدل داده عمومی است که الگوریتمهای رگرسیون می توانند در هر یک از مکانهای مشاهده شده (OL) اعمال شوند. هدف اصلی استفاده از یک مدل عمومی این است که آزمایش بهترین مدل مناسب برای همه OL ها عملی نیست. با تجزیه و تحلیل ویژگی ها از طریق تجزیه و تحلیل مؤلفه های اصلی ، هیستوگرام ها و ماتریس های همبستگی امکان درک کلی بهتر از داده ها را فراهم می کنند و یک مدل داده ای را ایجاد می کنند که هر مجموعه داده جداگانه می تواند از آن استفاده کند.

۴-۶-۱ بررسی ترافیک

در این بخش ، مشاهده ترافیک به تفصیل توزیع مقادیر ، تجمع داده ها ، داده های متنوع فصلی و متا داده های بیشتر با استفاده از تجسم بحث شده است.

پیچیدگی تجزیه و تحلیل مشاهدات در داده های بزرگ ، حجم اطلاعات است. درک برخی از جزئیات با استفاده از تجسم با نقشه های دقیق آسان تر است. لیست درمجموع ۴۷ مسیر در سراسر شبکه راه شهری است. یک مسیر می تواند تا ۲۵ پیوند مشاهده شده در دو جهت داشته باشد.

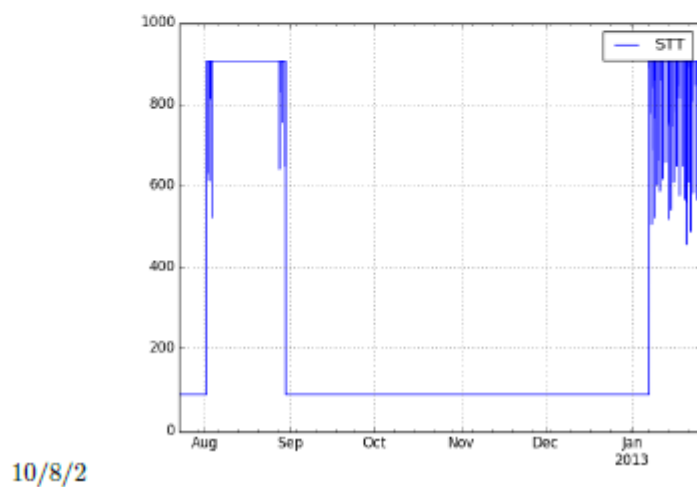
	direction	links	routes
2 count	698	698	698
min	1	1	1
4 max	2	25	47

در جدول ۴-۵ توزیع مقادیر از ۲۰۱۲/۰۷/۲۳ تا ۲۰۱۴/۰۴/۱۹ ۲۳:۵۰ نشان داده شده است.

جدول تعداد نمونه مشاهده شده مورد استفاده شمارش ، انحراف معیار استاندارد ، حداقل و حداکثر مقدار و توزیع کوانتی را نشان می دهد. توزیع کمی در ۰/۵۰ (٪۵۰) مقدار متوسط و ۰/۸۰ (٪۸۰) میانگین میانگین ۲۰٪ بالای مقادیر است. ۱۲۸ از ۶۹۸ مکان مشاهده شده دارای تعداد ۲۶۸۳۴ است. مکانهای مشاهده شده برای تجزیه و تحلیل نامعتبر هستند. حجم داده های از دست رفته برای انجام هرگونه تعویض مقادیر گم شده بسیار بزرگ است. برخی از مکانهای مشاهده شده با مقادیر شمارش ۲۶۸۳۴ مشخص شده اند که مکان های تکراری مکان دیگر را با یک تعداد مشاهده کامل تر نشان می دهند. در نتیجه تعداد مجموعه

داده های مشاهده کامل ۵۷۸ است. در شکل ۳-۵ نمایی از مشاهدات زمان سفر برای مکان ۲/۸/۱۰ را نشان می دهد.

<i>id</i>	<i>count</i>	<i>std</i>	<i>min</i>	<i>max</i>	<i>.20</i>	<i>.40</i>	<i>.60</i>	<i>.80</i>
1/1/1	91584	62	117	1045	118	127	133	159
1/1/2	91584	8	59	411	60	61	62	63
1/10/1	91584	51	7	773	7	7	7	99
1/10/2	91584	29	7	482	15	19	29	53
1/11/1	91584	22	18	242	18	18	18	27
1/11/2	91584	24	18	242	18	18	18	31
1/12/1	91584	11	9	130	10	10	11	16
1/12/2	91584	17	9	103	9	9	18	33
1/13/1	91584	17	5	263	12	23	35	39
1/13/2	91584	17	5	263	5	5	5	29
1/14/1	26834	2	5	86	10	11	11	11



شکل ۴-۵: مشاهدات زمان سفر

جدول ۴-۶: توزیع مقادیر

فصل پنجم

پیاده سازی

مقدمه

در این فصل ، مدل های داده برای پیش بینی را بر اساس ویژگی های مشاهده زمان سفر ، همسایگان مکانی و داده های آب و هوا ارائه خواهیم داد. ویژگی های مورد استفاده با تجزیه و تحلیل و کاهش ویژگی های ضروری ARIMA برای ساختن یک مدل عمومی تعیین می شود که در کلیه مکان های OL مشاهده شده مناسب خواهد بود. با استفاده از تکنیک های داده کاوی و تجسم ، یک مدل کلی ایجاد شده است.

۱-۵ انتخاب مدل زمان سفر استاندارد (STT)

در این بخش تمرکز بیشتر روی پیش بینی اوج زمان در روزهای هفته است. سه نوع نوسانات ترافیکی در بخش اکتشاف داده ها شناسایی شده و تجزیه و تحلیل توزیع مقادیر مشاهده را طبقه بندی می شوند. این اتصالات دسته های کم، متوسط تا زیاد را پوشش می دهد که در جدول ۱-۵ مشاهده می کنیم.

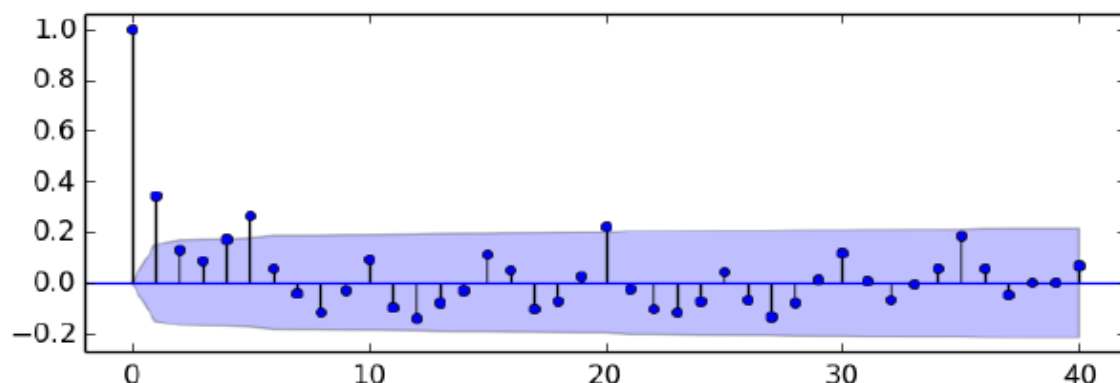
	Lag -1	Lag -2	Lag -3	Lag -4	Lag -5
≥ 0.5	50%	39%	34%	29%	30%
>0 and <0.5	49%	59%	63%	68%	69%
<0	0.2%	1.0%	1.5%	2.0%	1.0%

جدول ۱-۵: نوسانات ترافیکی

هر کدام از لگ ها (lag) نشان دهنده روز ترافیکی هستند که برای روزهای کاری در نظر گرفته شده اند. برای همبستگی مبتنی بر رابطه بین ماتریس ضریب همبستگی ، "P" و ماتریس کواریانس "C" از رابطه زیر استفاده گردید.

$$P_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} * C_{jj}}}$$




نتایج نشان می دهد که وجود رابطه در تاخیر ۱- و وخامت تدریجی با میزان فاصله تاخیر بیشتر می شود. در بعضی موارد ، همبستگی تاخیر ۴- که یک هفته از نظر روزهای کاری است ، یک افزایش جزئی مشاهده می شود.

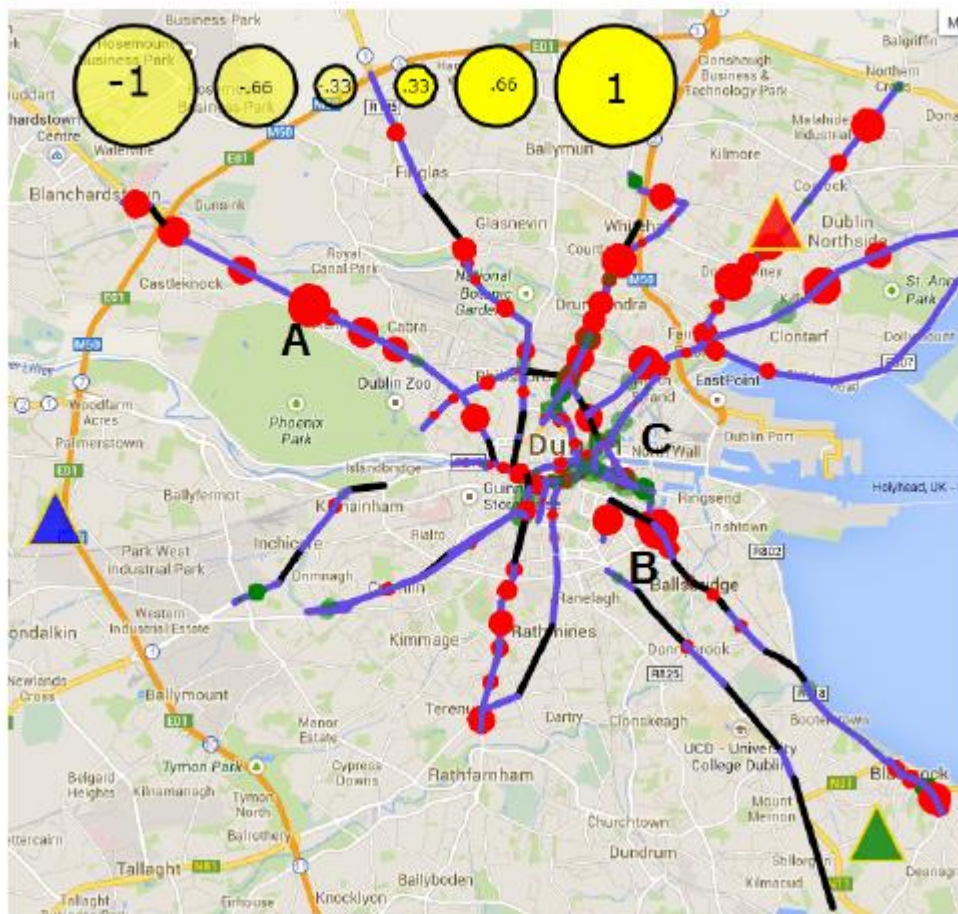


شکل ۵-۱: نمودار همبستگی روزانه

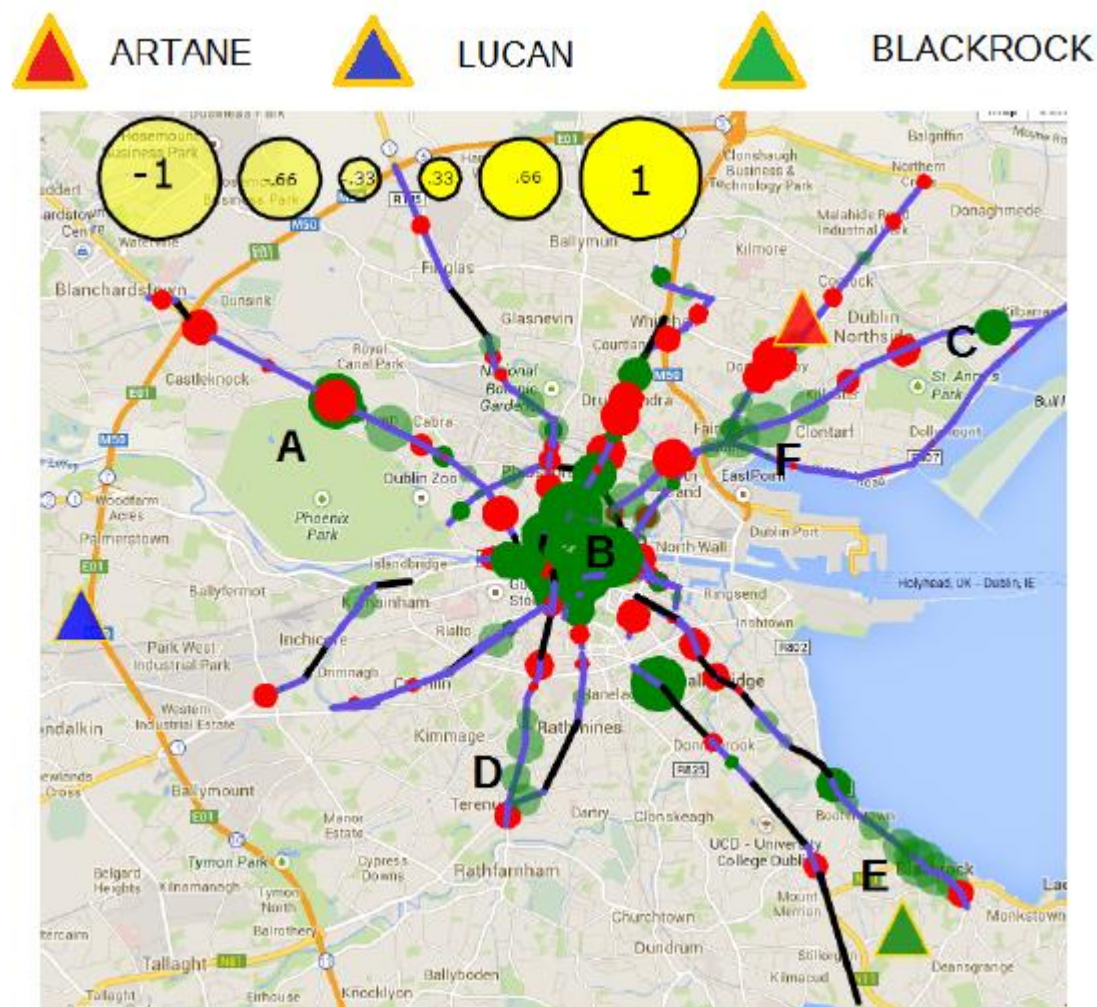
۲-۵ مدل آب و هوا

در این بخش ارتباط STT با شرایط آب و هوایی بارانی و دما مشاهده می شود. بخش اکتشاف این خصوصیات را به عنوان متغیرهای مناسب برای پیش بینی هوا شناسایی کرده است. همان معیار همبستگی برای آب و هوا مورد استفاده در بخش مدل سازی STT که ماتریس ضریب همبستگی و ماتریس کواریانس است. شکل [۲-۵ و ۳-۵] تصاویر با استفاده از نقشه های گوگل از همبستگی میزان بارش و آمار تجسم ارتباط همبستگی دما است. دایره های تجسم در نقشه قدرت همبستگی را با اندازه دایره و رنگ نمایانگر ایستگاه هواشناسی در آن قویترین همبستگی را نیز نشان می دهد. حلقه ها روی مکان مشاهده شده مربوط به آن قرار دارند. همبستگی منفی در نقشه هنگامی آشکار می شود که دایره آن دارای شفافیت ۰/۰ باشد و آن را نیمه شفاف می کند. مقادیر دامنه همبستگی بین ۱- تا ۱ است. در نقشه دایره ها ایستگاه را با مقدار همبستگی دورتر از ۰ نشان می دهد که می تواند منفی یا مثبت باشد. حلقه های زرد مقیاس و شفافیت را به عنوان نمونه ای از نمایش ارزش ارائه می دهند.

 ARTANE
  LUCAN
  BLACKROCK



شکل ۵-۲: همبستگی زمانهای اوج ترافیک و بارندگی



شکل ۵-۴: همبستگی زمانهای اوج ترافیک و دما

در شکل ۵-۴ قویترین همبستگی دما برای بارهای ورودی اوج بین ایستگاه آب و هوایی گسترش یافته است. هنگامی که درجه حرارت بالا و گرم است، نقشه به نظر می رسد در پارک های ملی Phoenix Park و St. Annes با برچسب A، C همبستگی مثبت داشته باشد. همچنین مرکز شهر تأثیر زیادی دارد. نشان می دهد که وقتی درجه حرارت گرم است، حجم ترافیک در پارک های کشور و مرکز شهر افزایش می یابد. مناطق حومه شهر از طرف دیگر ترافیک هنگام سرد شدن هوا افزایش می یابد همانطور که در

نقشه در D, E, F نشان داده شده است. و نشان می دهد که مردم به احتمال زیاد از اتومبیل های خود به عنوان حمل و نقل در هوای سرد استفاده می کنند.

در جدول ۵-۲ همبستگی بین سه نقطه مورد مطالعه و بارندگی و دما در زمانهای اوج نشان داده شده است.

Range	Lucan		Blackrock		Artane	
	Rain	Temp	Rain	Temp	Rain	Temp
$\geq 1 \ \& \ \geq 0.66$	0%	0%	0%	0%	0%	0%
$\geq 0.33 \ \& \ \leq 0.66$	5%	0%	7%	2%	8%	8%
$\geq 0.0 \ \& \ \leq 0.33$	71%	63%	40%	58%	65%	59%
$\leq 0.0 \ \& \ \geq -0.33$	28%	31%	60%	31%	34%	31%
$\leq -0.33 \ \& \ \geq -0.66$	0%	1%	3%	0%	2%	8%
$\geq -1 \ \& \ \leq -0.66$	0%	0%	0%	0%	0%	0%

جدول ۵-۲: همبستگی بارندگی و دما با ترافیک

در جدول ۵-۲ ارتباط متغیرهای آب و هوا با مکان مشاهده شده در محدوده کم بین ۰,۳۳- تا ۰,۳۳+ است. این نشان می دهد که تأثیر چندانی آب و هوا به مکانها ندارد. هر مکانی پایین تر از ۰,۳۳- یا بالاتر از ۰,۳۳+ دلالت بر بررسی بیشتر در مورد کیفیت جاده یا طراحی شبکه راه ممکن است نیاز به بررسی بیشتر داشته باشد.

۵-۳ اتصالات مدل پیش بینی

در این بخش الگوریتم های پیش بینی به مجموعه داده های متشکل از ویژگی های مورد بحث در انتخاب استاندارد زمان سفر، انتخاب مدل آب و هوا، انتخاب مدل فضایی اعمال می شود. هدف از این بخش تولید بهترین مدل مناسب برای تخمین STT روز بعد است. در این توضیحی از مجموعه داده ها ارائه می شود، برای برآورد بهترین مدل مناسب، مکانیزم امتیاز دهی برای ایجاد مقایسه الگوریتم برای هر مکان مشاهده شده مورد نیاز است.

۵-۳-۱ مجموعه داده های پیش بینی شده

متغیرهای داده مکانی و متغیرهای داده هواشناسی برخی از نویزهای مرتبط با مقدار STT را تشکیل می دهند. هر ۵۶۳ مکان مشاهده شده یک مجموعه داده منحصر به فرد است که از همان ویژگی های شرح داده شده در تشکیل شده است.

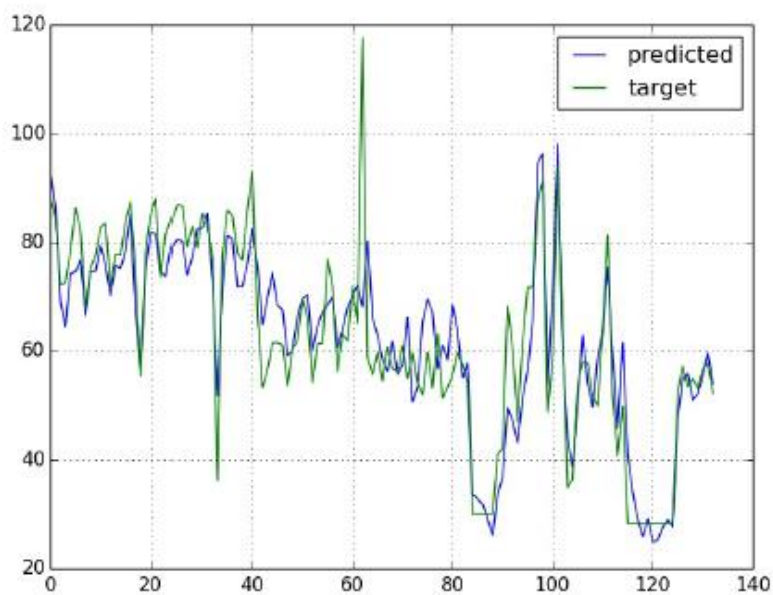
برای جلوگیری از بیش از اعتبار سنجی متقابل، از مجموعه داده ها به مجموعه داده های آموزش و آزمایش تقسیم می شود. اندازه داده آزمایش روی ۳۰٪ از مجموعه داده های کامل تنظیم شد. مجموعه داده شامل ۱۳۸ نمونه است.

۵-۳-۲ نتایج پیش بینی

برای ساختن یک مدل عمومی که سطح خوبی از دقت را برای هر منطقه OL مشاهده کرده مطابقت داشته باشد از الگوریتمهای زیر استفاده کردیم.

• رگرسیون خطی

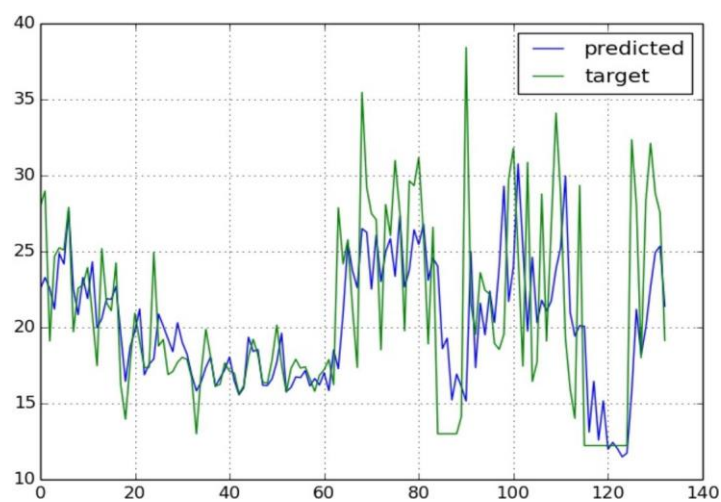
رگرسیون معمولی خطی بهترین الگوریتم پیش بینی را در اکثر موارد ارائه می کند. که دلیل آن توزیع استاندارد (STD) مقادیر دارای مقدار کمتر از ۱ است. نتایج این الگوریتم در شکل ۵-۵ قابل مشاهده است.



شکل ۵-۵: نتایج رگرسیون خطی

• رگرسیون برداری پشتیبان

(SVR) در صورت خطی بودن هسته می تواند مورد استفاده قرار گیرد. این الگوریتم همچنین عملکرد خوبی دارد. SVM نوعی الگوریتم ماشین بردار پشتیبانی است که می تواند برای مدل های خطی مناسب باشد.



شکل ۵-۶: نتایج رگرسیون برداری پشتیبان

۴-۵ مدل سازی ترافیک توییت

هدف از این بخش تجزیه و تحلیل رویکرد استفاده از توییت های دامنه ترافیک برای استخراج توییت از داده های زمان واقعی است که مربوط به حوزه ترافیک است.

Result

1	\label{algorithm_tweetresult}				
	precision	recall	f1-score	support	samples
3	Traffic	1.00	0.31	0.48	5000
5	Non-Traffic	0.59	1.00	0.74	5000
7	avg / total	0.80	0.66	0.61	10000

جدول ۳-۵: نتایج توییت

طبقه بندی کننده Agive Agresive Passive نمونه ای از یکی از الگوریتم های مورد استفاده برای طبقه بندی توییت های ترافیک در زمان واقعی است. در جدول ۳-۵ نتیجه توییت های زمان واقعی طبقه بندی شده به عنوان ترافیک نشان داده شده است.

فصل ششم

نتیجه گیری

مقدمه

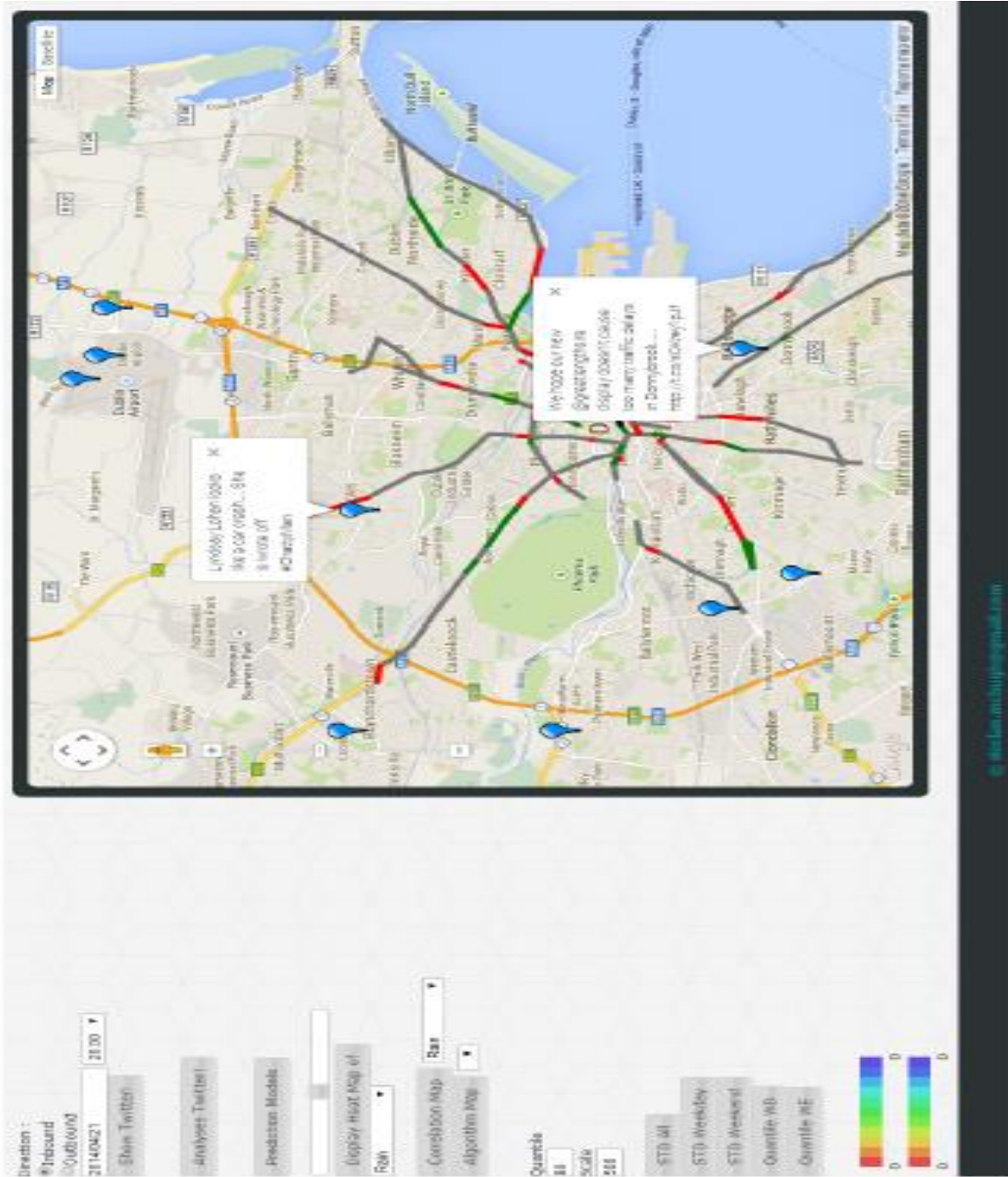
در نتیجه استفاده از رویکرد پیشنهادی برای غلبه بر چالش های ۴ V، رویکرد حجم داده هایی را ذخیره می کند که در یک سیستم واحد از مولفه های جدول زیر تهیه شده است.

<i>Data Source</i>	<i>Items</i>	<i>No. of Documents</i>
Traffic Observations	501,402,840	8,356,714
Real-time Tweets	3,048,310	116
User Tweets	5,267	5,267
Weather Records	229,311	2,103

جدول ۶-۱

۶-۱ پیاده سازی

Google Maps، JQuery و Python's چارچوب وب Django به شدت برای ایجاد برنامه ای استفاده شده است که نه تنها برای اکتشاف داده های مکانی بلکه در تحلیل الگوهای ترافیکی مانند نوسانات، تجزیه و تحلیل وقایع گذشته و پیش بینی ترافیک استفاده شده است. برنامه نتیجه عملکردی برای تجزیه و تحلیل بصری در، نوسانات جاده ها، تأثیر آب و هوا در مکان ها با توجه به زمان سفر و انجام تجزیه و تحلیل در مکان ها برای یک زمان خاص فراهم می کند.



جدول ٦-٢

۶-۲ تحلیل ترافیک

اندازه گیری تمام جنبه های فصلی به دلیل فقدان داده های با کیفیت به عنوان یک چالش به وجود آمده است. این یک مشکل رایج هنگام برخورد با داده های باز است. تجزیه و تحلیل روند سالانه ، سه ماهه و ماهانه نمی تواند انجام شود. این کار بیشتر در گرایشهای هفتگی و روزانه متمرکز بود. تفاوت اوج ترافیک روزهای آخر هفته و روزهای هفته واضح بود. اوج پایان هفته معمولاً در اواسط روز و هفته اوج روز در اوقات صبح یا عصر از ساعت ۲ بعد از ظهر بود. از طرفی بعلت عدم دسترسی به داده های شهر تهران از داده های خارجی در دسترس استفاده گردید.

روند در الگوهای ترافیکی بر اساس آب و هوا مشخص شد. نشان داده شد که افراد با درجه حرارت بالا به احتمال زیاد به مرکز شهر و پارک های عمومی سفر می کنند زیرا وقتی این اتفاق می افتد زمان مسافرت افزایش می یابد. تأثیر بارندگی با افزایش زمان سفر در نزدیکی روستاهای کوچک در Castleknock, Raheny, Drumcondra در میان دیگران بود. این می تواند نشانگر این باشد که روستا نمی تواند از افزایش گردش ترافیک برخوردار باشد ، در حالی که مردم احتمالاً بیش از پیاده روی در باران به سمت خرید خود رانندگی می کنند. این ممکن است نشانگر این باشد که مردم در هنگام باران مسافت های کوتاه را به یک روستای محلی سفر می کنند اما وقتی هوا خشک و گرم است ممکن است سفر کنند تا آنجا را در مرکز شهر بخرند.

۶-۳ مدل پیش بینی

مدل پیش بینی نهایی ترکیبی از SARIMA و Multivariate ARIMA شد. در این کار ممکن است به نظر برسد که الگوریتم های خطی بهتر از الگوریتم های غیر خطی از لحاظ دقت در پیش بینی باشند. همچنین باید در نظر گرفت که مدل داده طراحی شده یک مدل کلی است که همه مکانهای مشاهده شده را متناسب

می‌کند. یکی از محدودیت‌ها این بود که یک شبکه عصبی مصنوعی به عنوان بخشی از Python SciPy Toolkal قابل اجرا نیست. بنابراین تنها یک مدل غیرخطی با چهار الگوریتم خطی مقایسه شد. با توجه به تعداد نمونه‌های کلی آزمایش شده ممکن است موردی وجود داشته باشد که بیش از حد مناسب بر نتایج تأثیر بگذارد. بعضی از الگوریتم‌ها به دلیل عدم نوسان در انحراف استاندارد، از آن بهره‌مند شدند.

۶-۴ داشبورد تجزیه و تحلیل

داشبورد حاوی برخی از موارد مثبت کاذب برای زمان معین بود. رویکرد به دست آوردن توییت از یک دامنه خاص برای طبقه‌بندی داده‌های زمان واقعی امکان‌پذیر است. این دسته‌بندی توییتها با استفاده از مدل‌های طبقه‌بندی هوشمندتر مورد نیاز است.

۶-۵ کار آینده

برای آینده با توجه به کیفیت داده‌ها، تعداد نمونه‌ها محدود بود. مقایسه فصلی برای تعطیلات مدارس در طول تعطیلات تابستان یا زمستان مورد آزمایش قرار نگرفت. در تجسم نتایج پیش‌بینی، مکانهای مشاهده شده هیچ نوع شاخص عملکرد اصلی (KPI) را اعمال نمی‌کنند. وضعیت فعلی از رنگ آمیزی پیش‌بینی قرمز یا سبز بالاتر یا پایین‌تر از آنچه پیش‌بینی شده است استفاده می‌کند. روشی برای استفاده از محدوده رنگی به خواننده اجازه می‌دهد مقیاس پیش‌بینی را از نتیجه واقعی منحرف کند.

ارزیابی بیشتر در مورد الگوریتم‌ها، مکانیسم‌های نشانه‌گذاری و امتیازدهی به طبقه‌بندی ترافیک توییت‌ها برای بهبود کیفیت نتیجه ضروری است. فن‌آوری‌های استخراج متن مانند توقف لغت و بخشی از گفتار احتمالاً به بهبود طبقه‌بندی کمک می‌کند.

مراجع

- [1] Duckwon Chung et al. Road traffic big data collision analysis processing framework. *IEEE*, 2013.
- [2] Vinay Gavirangaswamy et al. Assessment of arima-based prediction techniques for road-traffic volume. 2013.
- [3] DubLinked. Trips data, 2012-2014. URL <http://www.dublinked.ie/datastore/datasets/dataset-215.php>.
- [4] McCreddie et al. Scalable distributed event detection for twitter. *IEEE*, 2013.
- [5] MongoDB. Big data explained @ONLINE, 2014. URL <http://www.mongodb.com/big-data-explained>.
- [6] Kalman R. E. A new approach to linear filtering and prediction problems. *IEEE*, 1960.
- [7] Dehuai Zeng et al. Short term traffic flow prediction using hybrid arima and ann models. 2008.
- [8] Wei-Chiang Hong et al. Seasonal adjustment in a svr with chaotic simulated annealing algorithm traffic flow forecasting model. 2010.

- [9] Kevin Keay and Ian Simmonds. The association of rainfall and other weather variables with road traffic volume in melbourne australia. 2004.
- [10] Bei Pan et al. Crowd sensing of traffic anomalies based on human mobility and social media. *IEEE*, 2013.
- [11] Stephen Dunne and Bidisha Ghosh. Weather adaptive traffic prediction using neurowavelet models. 2013.
- [12] Yorgos J. Stephanedes. Dynamic prediction of traffic volume through kalman filtering theory rescue. 1983.
- [13] Bowu Zhang et al. Traffic clustering and online traffic prediction in vehicle networks. *A Social Influence Perspective*, 2012.
- [14] Yousef-Awwad Daraghmi et al. Space-time multivariate negative binomial regression for urban short-term traffic volume prediction. *IEEE*, 2012.
- [15] Sri Krisna Endarnoto et al. Traffic condition information extraction & visualization from social media twitter for android mobile application. *IEEE*, 2011.
- [16] Amit Sheth. Transforming big data into smart data: Deriving value via harnessing volume, variety, and velocity using semantic techniques and technologies. 2014.

- [17] Shweta Pandey and Dr.Vrinda Tokekar. Prominence of mapreduce in big data processing. 2014.
- [18] Howard Gobioff Sanjay Ghemawat and Shun-Tak Leung. The google file system, October, 2003.
- [19] Duckwon Chung et al. United nations global pulse, 2012, big data for development: Challenges & opportunities. May 2012.
- [20] Executive Office of the President. Office of science and technology policy. *IEEE*, 2012.
- [21] Brito et al. Scalable and low-latency data processing with streammapreduce. *IEEE*, 2011.
- [22] Pricilla Hancock Kristopher Reese, Russell Bessette. Knowyourcolors: Visual dashboards for blood metrics and healthcare analytics. *IEEE*, 2013.
- [23] Carlos Costa Lu'is Bastiao Silva Louis Beroud and Jos'e Luis Oliveira. Medical imaging archiving: a comparison between several nosql solutions. 2013.
- [24] Dr. Vincent Granville. Developing analytic talent: Becoming a data scientist. *Wiley*; *1st edition*, 2014.
- [25] Gong Jun et al. Forecasting urban traffic flow by svr. 2013.

[26] Machine Learning in Python Scikit-learn. Scikit-learn developers, 2010 - 2014. URL <http://scikit-learn.org/>.

[27] Wunderground. Wunderground, 2012 - 2014. URL <http://www.wunderground.com>.

[28] Daniel J Tulloch. A garch analysis of the determinants of increased volatility of returns in the european energy utilities sector since liberalisation. *IEEE*, 2012.

[29] Rohit Dhawan Sven F. Crone. Forecasting seasonal time series with neural networks: A sensitivity analysis of architecture parameters. *IEEE*, 2007.

[30] et al Koby Crammer. Online passive-aggressive algorithms. 2006.

[31] Rahul Khokale#1, Ashwini Ghate "Data Mining for Traffic Prediction and Analysis using Big Data" International Journal of Engineering Trends and Technology (IJETT) – Volume 48 Number 3 June 2017 ISSN: 2231-5381 <http://www.ijettjournal.org> Page 152

[32] Yongmei Zhaoa,b, Hongmei Zhangb, Li Anb, Quan Liu "Improving the approaches of traffic demand forecasting in the big data era" <https://doi.org/10.1016/j.cities.2018.04.015> April 2018; 2018

[33] حاجی حسینلو منصور، شریفیان میثم، پاک روشن بیژن "تحلیل ترافیکی شبکه های درون شهری با استفاده از نرم افزار شبیه سازی) Corsim مورد مطالعه: خیابان شهید دستغیب شهر تهران) " مطالعات مدیریت ترافیک: بهار ۱۳۸۹، دوره ۵، شماره ۱۶؛ از صفحه ۱۱ تا صفحه ۲۴.

[34] رضایی, نرگس و امیر شکیبامنش, ۱۳۹۷, تحلیل حوادث ترافیک شهری با کمک نرم افزار SANET (نمونه مورد مطالعه: منطقه ۸ شهر تبریز), سومین کنفرانس بین المللی عمران, معماری و طراحی شهری, تبریز, دبیرخانه دائمی کنفرانس -دانشگاه میعاد با همکاری دانشگاه هنر اسلامی تبریز-دانشگاه خوارزمی- دانشگاه شهرکرد, https://www.civilica.com/Paper-ICCACS03-ICCACS03_086.html

[35] عسکری, مریم و بهروز مینایی بیدگلی, ۱۳۹۴, تحلیل ترافیک تاکسی های شهری به کمک تکنیک های داده کاوی, کنفرانس بین المللی یافته های نوین پژوهشی درمهندسی برق و علوم کامپیوتر, تهران, موسسه آموزش عالی نیکان, https://www.civilica.com/Paper-COMCONF01-COMCONF01_125.html

واژه نامه

واژه نامه انگلیسی به فارسی

Index Chunked

فهرست قطع شده

Big Data 4 V

حجم، سرعت، صحت و تنوع

Big Data

داده های بزرگ

Autoregressive Integrated

میانگین متحرک خود پیش بینی

Map Reduction

کاهش نقشه

Split Key Partitioning

بخش بند از طریق تقسیم کلید

lag

تاخیر

Abstract

Smart City is an innovative work that has been done in many European cities. The goals of smart cities are to improve planning, social management and infrastructure management of the city. Cities like Dublin, Lyon, Amsterdam and Barcelona are examples of smart cities. Many European cities, including Dublin, have an active intelligence program that protects the privacy of many data portals open to the public in such cities for analytics, and providing an intuitive dashboard for It is important to analyze it. Because traditional methods are not capable of predicting good quality, research is conducted to identify traditional methods of traffic forecasting and analysis in area forecasting and spatial data mining. Researchers blame data analytics for identifying problems and developing them. Further research will focus on the concept of how traditional traffic forecasting approaches are analyzed with historical data and social media. The CRISP DM data mining method will be used throughout the project. This method is an industry standard for data mining. This method plays an important role in the creation of techniques and tools. In general, our method is based on the use of data in standard datasets that will be analytically processed after the algorithm is pre-processed and applied.

Keywords: Traffic Forecasting, Traffic Pattern Analysis, Data Mining



Payam Noor University

Faculty of Engineering

Thesis

Master of Science Degree in Computer Engineering (Software)

Department of Computer Engineering and Information Technology

Traffic prediction using evolutionary algorithms and big data

Shahrooz bazrafshan

Supervisor:

Dr.Seyyed Ali Razavi Ebrahimeia

June 2017

