نشریه علمی پژوهشی مهندسی فناوری اطلاعات مکانی

سال سوم، شماره چهارم، زمستان ۱۳۹۴

Vol.3, No.2, Winter 2016

۸۳-۹۵

الگوریتمی برای فشرده‌سازی خطوط سیر مکانی با حفظ ماهیت معنایی

سمیه عاقل شاه‌نشین [۱]، سیمین سادات میروهابی [۱]، رحیم علی‌عباسپور [۲*]

۱- دانشجوی کارشناسی ارشد سیستم‌های اطلاعات مکانی ـ دانشکده مهندسی نقشه برداری و اطلاعات مکانی ـ پردیس دانشکده های فنی ـ دانشگاه تهران

۲- استادیار دانشکده مهندسی نقشه برداری و اطلاعات مکانی ـ پردیس دانشکده های فنی ـ دانشگاه تهران

تاریخ دریافت مقاله: ۹۴/۰۹/۱۲ تاریخ پذیرش مقاله: ۹۴/۱۲/۱۰

چکیده

یک راه معمول برای ذخیره اطلاعات مکانی-زمانی اشیاء در حال حرکت، نمایش مسیر حرکت شئ متحرک به شکل خط سیر سه‌بعدی (موقعیت جغرافیایی نقاط به همراه زمان) است. در سال‌های اخیر تحقیقات گسترده‌ای در حوزه خط سیر انجام شده است. با این حال، در این مطالعات، ایده خط سیر معنایی یک مفهوم نسبتاً جدید است که با هدف انجام آنالیزهای معنایی مؤثر روی داده‌ها انجام می‌شود. در خط سیر معنایی که یک نمایش ثانویه از خط سیر مکانی است، حرکت شئ به شکل دنباله‌ای از توقف‌ها و حرکت‌ها توصیف می‌شود. تولید خط سیر معنایی از داده‌های خام جمع‌آوری شده، یک فرآیند با چندین گام پردازش است که با توجه به حجم عظیم داده‌ها، یکی از پیش‌پردازش‌های موردنیاز کاهش تعداد نقاط خط سیر با حفظ دقت موردنیاز با استفاده از تکنیک‌های فشرده‌سازی است. با وجود این، اغلب تکنیک‌های کاهش داده خط سیر که بر اساس ساده‌سازی خطی هستند، قادر به حفظ مناطق توقف و حرکت نیستند. در این مقاله روشی برای فشرده‌سازی داده‌های خط سیر بر اساس سرعت نقاط ارائه شده است که از ترکیب دو تابع فاصله در درونیابی و محاسبه خطای نقاط استفاده کرده است. تابع فاصله اول بر اساس سرعت نقاط است که برای محاسبه خطای تقریب خط سیر استفاده شده و تابع فاصله دوم تابعی بر اساس توسعه الگوریتم شناخته شده داگلاس-پوکر است که از فرض ثابت بودن شتاب در محاسبه خطا در تقریب استفاده کرده است. الگوریتم ارائه شده روی داده‌های واقعی خط سیر پیاده‌سازی شده و نتایج به دست آمده حاکی از بهبود عمکرد در حفظ مناطق توقف در مقایسه با الگوریتم‌های فشرده‌سازی دیگر است.

کلیدواژه‌ها: خط سیر معنایی، فشرده‌سازی، مدل توقف-حرکت، E_1-E_2

نویسنده مکاتبه‌کننده: تهران-خیابان کارگر شمالی، بعد از تقاطع آل احمد، پردیس دانشکده‌های فنی دانشگاه تهران، دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی

تلفن: ۸۸۰۰۸۸۴۱

Email: abaspour@ut.ac.ir

۱- مقدمه

امروزه تعیین موقعیت توسط دستگاه‌های سیار مجهز به GPS افزایش چشمگیری داشته است. به این ترتیب حجم عظیمی از اطلاعات مکانی به‌طور مداوم و روزانه تولید می‌شود. تعیین موقعیت اشیاء متحرک در خدمات مکان‌مبنا کاربردهای گسترده‌ای از قبیل مدیریت و برنامه‌ریزی شهری، مدیریت حمل و نقل و ترافیک، گردشگری، الگوکاوی مسیر مهاجرت حیوانات و غیره دارد. با این حال، اغلب این خدمات روی موقعیت فعلی کاربران تمرکز می‌کنند و اطلاعاتی که می‌تواند از تاریخچه حرکت کاربران به دست آید را کمتر موردتوجه قرار می‌دهند. با این حال، غیر از اطلاعات خامی که مستقیما از دستگاه‌های تعیین موقعیت به دست می‌آید، خدمات و کاربردهای نسل‌های بعد نیاز به این دارند تا منطق رفتاری و انگیزه حرکت را نیز بررسی کنند تا اطلاعات بیشتری را فراهم کنند. خط سیر [1] مدلی برای نمایش مسیر حرکت جسم متحرک است که معمولاً به صورت رشته‌ای از موقعیت‌های زمان‌دار به شکل (P_i, t_i) تعریف می‌شود که در آن P_i موقعیت دوبعدی نقطه یعنی (x_i, y_i) است و t_i نیز زمان ثبت مربوط به نقطه است [۱].

در حوزه مطالعات مربوط به خط سیر، خط سیر معنایی[2] مفهوم نسبتاً جدیدی است که با هدف انجام آنالیزهای معنایی مؤثر روی داده‌ها ایجاد می‌شود. خط سیر معنایی یک نمایش ثانویه از مسیر حرکت شئ است که فهم مسیر را ساده‌تر می‌کند. ایده اصلی این است که مسیر حرکت شئ دنباله‌ای از بخش‌های همگن و معنی‌دار است. یک مدل مفهومی رایج در شناسایی این بخش‌های همگن، مدل توقف-حرکت است با این فرض اساسی که شئ متحرک برای انجام فعالیت‌های مهم و یا در مکان‌های مهم توقف می‌کند و یا سرعت خود را کاهش می‌دهد. به این ترتیب مناطق توقف معمولاً به صورت تراکم نقاط در مناطق توقف (به دلیل توقف شئ متحرک و یا کاهش سرعت) و یا نبود سیگنال GPS (به دلیل ورود شئ متحرک به فضای بسته) در خط سیر ظاهر می‌شود. از این رو می‌توان نتیجه گرفت، در حالی که در مباحث مربوط به خطوط سیر هندسی، مسیر حرکت شئ متحرک مبنای تحلیل داده‌های خط سیر است، در خطوط سیر معنایی، علاوه بر مسیر حرکت شئ متحرک، شناسایی و تحلیل مکان‌های توقف در خط سیر نیز از اهمیت ویژه‌ای برخوردار است.

شکل ۱ مثالی از خط سیر هندسی و خط سیر معنایی متناظر با آن را برای گردش یک روزه یک گردشگر نشان می‌دهد. در شکل ۱.(بالا)، خطوط سیر به شکل نقاط ساده بدون اطلاعات معنایی نشان داده شده‌اند. در شکل ۱.(پایین)، اطلاعات معنایی با خطوط سیر ترکیب شده‌اند. همانگونه که مشاهده می‌شود، در خط سیر معنایی، مکان‌های مهم از نظر کاربرد با اطلاعات معنایی غنی‌سازی شده‌اند. در این مثال، مکان‌های مهم شامل اماکن تفریحی و گردشگری و محل اقامت هستند.

نگهداری و ذخیره اطلاعات خط سیر در پایگاه داده اشیاء متحرک به سرعت می‌تواند به حجم عظیمی از داده‌ها تبدیل شود. از این رو ارائه راهکارهای مناسب برای مدیریت مؤثر این حجم عظیم از داده‌ها ضروری به نظر می‌رسد. از یک سو غنی‌سازی معنایی داده‌های خط سیر مکانی یک راه مؤثر در کاهش این حجم عظیم داده‌هاست. از سوی دیگر در پردازش داده‌ها که با هدف آماده‌سازی داده‌های خط سیر برای غنی‌سازی معنایی انجام می‌شود، استفاده از تکنیک‌های فشرده‌سازی برای حذف داده‌های اضافی می‌تواند مورد استفاده قرار گیرد. با این حال سؤال اساسی این است که می‌توان با حفظ دقت

[1] Trajectory
[2] Semantic Trajectory

موردنیاز به حذف تعدادی از داده‌ها پرداخت. با توجه به ماهیت خطی حرکت اشیای متحرک، بسیاری از کاربردهای مربوط به خدمات مکان‌مبنا دارای الگوهای قابل پیش‌بینی هستند؛ در نتیجه بسیاری از اطلاعات اضافی می‌توانند از خط سیر حذف شوند.

شکل ۱: خط سیر هندسی (بالا) که معادل خط سیر معنایی (پایین) است[۲]

اغلب تکنیک‌های کاهش داده خط سیر، بر اساس روش‌های ساده‌سازی خط هستند. در این روش‌ها هدف از فشرده‌سازی، کاهش تعداد نقاط با حفظ ماهیت هندسی خط سیر می‌باشد. با این حال در خطوط سیر معنایی حفظ ماهیت خط سیر، که مستلزم حفظ نقاط توقف می‌باشد، از اهمیت بالایی برخوردار است. همانطور که در شکل ۲.(۲) مشاهده می‌شود، یک تکنیک فشرده‌سازی بر اساس ساده‌سازی خطی قادر به حفظ مناطق توقف در خط سیر نیست. اما در شکل ۲.(۳) تکنیک فشرده‌سازی معنایی علاوه بر کاهش حجم داده‌های خط سیر، مناطق توقف را نیز حفظ کرده است.

شکل ۲- ضرورت فشرده‌سازی خطوط سیر برای حفظ مناطق توقف، (۱) خط سیر اصلی، (۲) خط سیر فشرده‌سازی‌شده با تکنیک ساده‌سازی خطی، (۳) خط سیر فشرده‌سازی با هدف حفظ مناطق توقف

به علت وجود کاستی‌های مذکور در روش‌های موجود، این مقاله سعی در ارائه روشی جدید برای رفع این مشکلات دارد. الگوریتم ارائه شده با در نظر گرفتن سرعت و زمان حرکت شئ متحرک و انجام پردازش‌های لازم روی این دو پارامتر قادر به فشرده‌سازی خط سیرهای مکانی با حفظ مناطق توقف و حرکت برای پردازش‌های بعدی در خط سیر است. نتیجه این تحقیق می‌تواند به عنوان مسیری جدید در فشرده‌سازی خطوط سیر معنایی برای استفاده در آنالیزهای مکانی و معنایی مورد استفاده قرار گیرد.

ساختار مقاله پیش‌رو به این شکل است: بخش ۲ به مرور مطالعات پیشین در زمینه فشرده‌سازی خطوط سیر هندسی و معنایی می‌پردازد. بخش ۳ به بیان مبانی نظری این تحقیق پرداخته است. در بخش ۴ روش پیشنهادی معرفی و تحلیل شده است و مراحل مختلف روش مورد بررسی قرار گرفته است. در بخش ۵ الگوریتم فشرده‌سازی ارائه شده روی داده‌های واقعی پیاده‌سازی شده و نتایج آن

مورد ارزیابی قرار گرفته است. بخش آخر نیز به جمع‌بندی و نتیجه‌گیری می‌پردازد.

۲- مروری بر مطالعات پیشین

مطالعات گسترده‌ای در حوزه فشرده‌سازی خط سیر انجام شده است. تکنیک‌های ارائه شده برای فشرده‌سازی خط سیر بر اساس نوع خط سیر را می‌توان به دو دسته تقسیم کرد:

۱. تکنیک‌های فشرده‌سازی مبتنی بر خطوط سیر هندسی

۲. تکنیک‌های فشرده‌سازی مبتنی بر خطوط سیر معنایی

در ادامه به بررسی و ارزیابی نمونه‌هایی در هر یک از دسته‌بندی‌های فوق پرداخته می‌شود.

۲-۱- تکنیک‌های فشرده‌سازی مبتنی بر خطوط سیر هندسی

به طور کلی مطالعات مربوط به تکنیک‌های کاهش داده خطوط سیر هندسی را می‌توان به دو بخش اصلی بر اساس نحوه بررسی و ورود داده‌ها در الگوریتم تقسیم کرد: فشرده‌سازی آفلاین و فشرده‌سازی آنلاین. در تکنیک‌های فشرده سازی آفلاین یا دسته‌ای[1] ابتدا یک مجموعه کامل از نمونه‌ها جمع‌آوری شده و سپس فشرده‌سازی مجموعه داده‌ها با حذف داده‌های اضافی انجام می‌شود. از آنجا که این تکنیک‌های فشرده‌سازی از کل مجموعه داده‌ها به صورت یکجا استفاده می‌کند، نتایج آنها منجر به دستیابی به دقتی بهتر نسبت به سایر تکنیک‌ها می‌شود. به عنوان مثال این تکنیک‌ها برای بارگذاری و تحلیل داده‌های خط سیر در وب‌سایت‌های مختلف چون EveryTrail4[2] و bikely5[3] (شکل ۳) بسیار مناسب هستند[۳]در ادامه به بررسی نمونه‌هایی از این تکنیک پرداخته می‌شود.

شکل۳- نمایی از خط سیر هندسی در سایت bikely5

شناخته شده‌ترین الگوریتم دسته‌ای، الگوریتم داگلاس-پوکر[4] [۴]است. این الگوریتم روشی برای ساده‌سازی خط با کاهش تعداد نقاط است. در این روش، اندازه‌گیری خطا بر اساس فاصله اقلیدسی قائم انجام می‌شود و از میانگین و یا مجموع فواصل خطا برای ارزیابی دقت تقریب استفاده می‌شود. این روش اندازه‌گیری خطا به شدت به تعداد نقاط نمونه وابسته است. ایده تصویر تمام نقاط ممکن در خط سیر اصلی روی پاره خط تقریبی نقاط، که اساس این روش است، تنها ویژگی‌های مکانی نقاط خط سیر را در نظر می‌گیرد و بعد زمانی نادیده گرفته می‌شود.

فاصله اقلیدسی همزمان، روشی برای اندازه‌گیری خطا برای نقاط است که در الگوریتم‌های کاهش داده خط سیر نیز می‌تواند استفاده شود. برای پرداختن به این موضوع مارتینا و دیبای[۵]با هدف اتخاذ معیارهای جدید خطا که نرخ فاصله زمانی[5] نامیده می‌شود، به جای فاصله اقلیدسی قائم از فاصله اقلیدسی همزما ن[6] استفاده کردند که معیار دقیق‌تری برای اندازه‌گیری خطا ارائه می‌دهد، چون هر دو ویژگی هندسی و زمانی شئ متحرک را در نظر می‌گیرد.

تابع خطا در الگوریتم داگلاس پوکر از نقاطی استفاده می‌کند که بیشترین انحراف را نسبت به خط سیر اصلی دارد. با این حال هیچ تضمینی وجود ندارد که این نقاط

[1] batched
[2] http://www.everytrail.com
[3] http://www.bikely.com

[4] Douglas-Peucker algorithm
[5] Time distance ratio
[6] Time synchronized

بهترین انتخاب باشند. الگوریتم بلمن[1][۶]از تکنیک برنامه نویسی پویا برای تقریب نقاط در فضای یک بعدی استفاده می‌کند. این الگوریتم می‌تواند برای تقریب خط سیر در فضای دو بعدی نیز تعمیم داده شود. با این حال، از انجا که این روش یک تابع پیوسته را تقریب می‌زند در کار با حلقه‌ها که در داده‌های خط سیر به وفور دیده می‌شود، دچار مشکل است. از سوی دیگر پیچیدگی زمانی بالای این روش، مشکل بعدی این الگوریتم است.

بسیاری از خدمات مکان مبنا نیاز به بروزرسانی آنلاین داده‌های خط سیر دارند. به عنوان مثال در مدیریت حمل و نقل و نظارت بر ترافیک از تکنیک‌های فشرده‌سازی آفلاین نمی‌توان به طور مستقیم استفاده کرد. الگوریتم نمونه برداری مخزن [2]در[۷]، برای پردازش داده‌های خط سیر یک الگوریتم مناسب است. در این روش موقعیت n نقطه ابتدایی به‌عنوان تقریب خط سیر ذخیره‌سازی می‌شوند. با اضافه شدن هر نقطه در این روش، الگوریتم با محاسبه یک تابع احتمالی حذف یا حفظ نقطه را مشخص می‌کند.

الگوریتم پنجره متحرک[3][۸]، که با هدف داده‌کاوی سری‌های زمانی ایجاد شده است برای تقریب خط سیر نیز می‌تواند مورد استفاده قرار گیرد. در[۵]، یک روش کاهش داده آنلاین بر اساس الگوریتم پنجره متحرک ارائه شده است. در روش پشنهادی از هر دو فاصله اقلیدسی قائم و اقلیدسی همزمان برای کاهش داده خط سیر استفاده شده است.

علاوه بر تکینک‌های کاهش داده آفلاین وآنلاین که از موقعیت نقاط خط سیر برای کاهش داده خط سیر استفاده می‌کنند، گروه دیگر از روش‌ها، از ویژگی‌های دیگر خط سیر نیز در تقریب استفاده می‌کنند. به عنوان مثال در[۹]، با ارائه یک روش پیش‌بینی از سرعت و جهت در تصمیم‌گیری برای حفظ یا حذف نقاط از مجموعه داده استفاده شده است. در الگوریتم پیشنهادی با تعریف منطقه امن به ارزیابی هر یک از نقاط خط سیر پرداخته شده است.

۲-۲- تکنیک‌های فشرده‌سازی مبتنی بر خطوط سیر معنایی

روش‌های کاهش داده خطوط سیر معنایی با هدف حفظ مناطق توقف و حرکت در مجموعه داده‌ها ارائه شده‌اند. در[۱۰]، یک روش کاهش داده آنلاین بر اساس حفظ نقاط مهم در خط سیر ارائه شده است. این روش روی داده‌های خط سیر کشتی پیاده سازی شده است. در روش پیشنهادی، موقعیت هر کشتی به‌صورت آنلاین به سرور ارسال می‌شود. سپس سرعت لحظه‌ای هر یک از نقاط به طور تقریبی با کمک نقطه قبل محاسبه می‌شود. در الگوریتم ارائه شده نقاط شاخص در خط سیر به چندین دسته تقسیم می‌شوند. سپس با تعریف سرعت هر یک از نقاط و مقایسه سرعت حاصل با حد آستانه تعریف شده برای هر گروه از نقاط شاخص تعریف شده، الگوریتم به ارزیابی نقاط می‌پردازد. اگر سرعت نقاط خط سیر در هیچ یک از گروه‌های تعریف شده نقاط شاخص نباشد، نقاط از مجموعه داده حذف می‌شوند.

در[۱۱]، دو الگوریتم KiT[4] و PaT[5] برای کاهش داده خط سیر ارائه شده است. در الگوریتم KiT نقاط مهم به ۵ دسته تقسیم می‌شود و نقطه‌ای که در این ۵ دسته قرار نگیرد از خط سیر حذف می‌شود. الگوریتم PaT نیز با خط سیر به صورت پلیگون برخورد می‌کند و ویژگی‌های حرکت با آنالیز ویژگی‌های هندسی پلیگون به دست آید.

در روش‌های ارائه شده در[۱۰,۱۱]شناسایی نقاط مهم

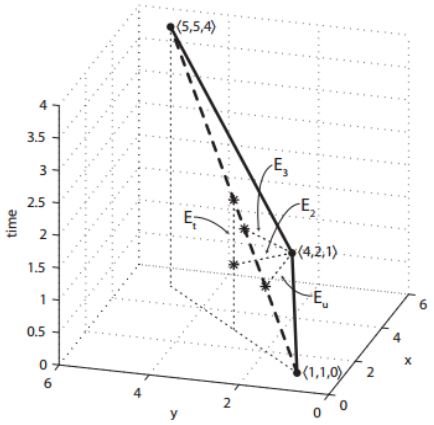
[1] Bellman
[2] reservoir
[3] Sliding window
[4] Key-in-Trajectory
[5] Polygon-as-Trajectory

در داده‌های خط سیر نیازمند معرفی بازه تعیین کننده هر یک از گروه‌هاست. با این حال، تعیین این بازه برای هر یک از گروه‌ها یک چالش است. از سوی دیگر، در این نوع روش‌های کاهش داده، مقادیر شناسایی شده در هر گروه به مقادیر حد بالا و حد پایین در هر بازه حساس هستند و با وجود کوچکترین اختلاف با مقادیر مرزی ممکن است بسیاری از نقاط به اشتباه از مجموعه داده‌ها کنار گذاشته شوند.

در این مقاله روش کاهش داده‌ای ارائه شده است که از تلفیق دو تابع در فشرده‌سازی استفاده می‌کند. تابع نخست یک تابع بر اساس سرعت نقاط است و تابع دوم نیز توسعه الگوریتم داگلاس-پوکر است. الگوریتم ارائه شده از دو تابع فاصله بازگشتی برای اندازه‌گیری خطای تقریب استفاده می‌کند. به این ترتیب روش ارائه شده مشکلات موجود در روش‌های[۱۰,۱۱]را که به دلیل دسته‌بندی نقاط و تعیین بازه تعیین کننده نقاط مهم ایجاد شده است، نخواهد داشت. در ادامه مقاله، پس از معرفی مفاهیم اولیه، روش پیشنهادی ارائه می‌گردد.

۳- مبانی نظری تحقیق

تکنیک‌های فشرده‌سازی استاندارد بر اساس ساده‌سازی خطی برای حفظ نقاط مهم در خط سیر کافی نیستند. در مباحث مربوط به خط سیر معنایی، با تمرکز بر مدل توقف- حرکت، نقاط مهم در خط سیر همان مناطق توقف هستند که در آن شئ متحرک در حداقل بازه زمانی تعریف شده بر اساس کاربرد موردنظر در آن منطقه باقی بماند. در این کار فشرده‌سازی بر اساس اصول PLS[1] توسعه یافته است که در آن برای حفظ مناطق توقف و حرکت، پارامتر سرعت نیز در طول خط سیر در نظر گرفته می‌شود. توابع خطاهای مختلفی می‌تواند در مدل PLS استفاده شود. شکل ۴ نمونه‌ای از توابع خطا را نشان می‌دهد، که در ادامه به مرور هر یک از توابع پرداخته می‌شود.



شکل۴- روابط بین توابع مختلف فاصله [12]

E_2: این تابع خطا همان تابع استفاده شده در الگوریتم داگلاس ـ پوکر است که در آن خط سیر به صورت یک خط در فضای دوبعدی تعریف می‌شود و از پارامتر زمان صرفنظر می‌شود.

در رابطه ۱، E2 فاصله قائم نقطه (x_i, y_i) را از (x_i', y_i') روی پاره خط واصل دو نقطه ابتدا و انتهای خط سیر یعنی (x_1, y_1) و (x_n, y_n) نشان می‌دهد. همانطور که مشاهده می‌شود استفاده از تابع فاصله E2 به تنهایی برای فشرده‌سازی خط سیر کافی نیست، چون بعد زمان در این تابع فاصله نادیده گرفته می‌شود.

E_3: در این تابع خطا، زمان به صورت بعد سوم مکانی تعریف می‌شود.

در رابطه ۲، E_3 فاصله نقطه (x_i, y_i, t_i) از پاره‌خط واصل دو نقطه ابتدا و انتهای خط سیر یعنی (x_i, y_i, t_i) و (x_n, y_n, t_n) در هر تقریب است.

E_u: در این تابع خطا، زمان به صورت بعدی است که به‌صورت متفاوت عمل می‌کند و درونیابی خطی آن بر اساس زمان روی پاره‌خط انجام می‌شود. این تابع فاصله همان فاصله اقلیدسی همزمان[2] نام دارد.

[1] Piecewise Linear Segmentation

[2] Synchronized Euclidian distance

شکل ۵ درونیابی خطی بر اساس سرعت ثابت را در فضای مکانی در تابع فاصله اقلیدسی همزمان نشان می‌دهد.

رابطه (۱) $$E_2\big((x_i,y_i),(x_1,y_1),(x_n,y_n)\big) = \|(x_i-x_i')^2+(y_i-y_i')^2\|$$

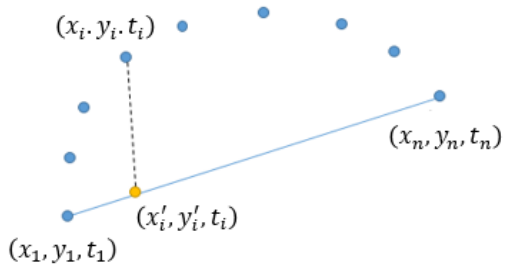
رابطه (۲) $$E_3\big((x_i,y_i,t_i),(x_1,y_1,t_1),(x_n,y_n,t_n)\big) = \|(x_i-x_i')^2+(y_i-y_i')^2+(t_i-t_i')^2\|$$

رابطه (۳) $$SED\big((x_i,y_i,t_i),(x_1,y_1,t_1),(x_n,y_n,t_n)\big) = \|(x_i-x_i')^2+(y_i-y_i')^2\|$$

رابطه (۴) $$x_i' = x_i + v_{1n}^x\cdot(t_i-t_1), \qquad y_i' = y_i + v_{1n}^y\cdot(t_i-t_1)$$

رابطه (۵) $$v_{1n}^x = \frac{x_n-x_1}{t_n-t_1},\ v_{1n}^y = \frac{y_n-y_1}{t_n-t_1}$$

روابط ۳ تا ۵ محاسبات تابع خطای اقلیدسی همزمان را نشان می‌دهند. همانطور که مشاهده می‌شود، نقطه (x_i',y_i',t_i)، نقطه همزمان متناظر با (x_i,y_i,t_i) روی پاره‌خط واصل بین ابتدا و انتهای خط سیر یعنی (x_1,y_1,t_1) و (x_n,y_n,t_n) در هر تقریب است. به این ترتیب تابع فاصله اقلیدسی همزمان از هر دو بعد مکانی و زمانی خط سیر در تقریب استفاده می‌کند.



شکل ۵- درونیابی خطی بر اساس سرعت ثابت در تابع فاصله اقلیدسی همزمان[۵]

E_t: این فاصله که کم و بیش دوگان تابع خطای E_u نیز هست، به جای تعیین اختلاف مکانی در زمان‌های یکسان (تعریف E_u) ، از اختلاف زمانی در مکان‌های یکسان یا مکان‌های نزدیک به هم استفاده می‌کند. فشرده‌سازی با این تابع خطا کامل نیست چون خطا در بعد مکانی نادیده گرفته می‌شود.

رابطه (۶) $$E_t\big((x_i,y_i,t_i),(x_1,y_1,t_1),(x_n,y_n,t_n)\big) = \sqrt{(t_i-t_i')}$$

در رابطه ۶، t_i' زمان نقطه (x_i',y_i')، روی نگاشت دوبعدی پاره خط واصل دو نقطه (x_1,y_1,t_1) و (x_n,y_n,t_n) روی صفحه xy است که نزدیک‌ترین نقطه به (x_i,y_i,t_i) است.

واضح است که استفاده از توابع فاصله E_2 و E_t برای فشرده‌سازی خط سیر کافی نیست، چون به ترتیب بعد زمانی و مکانی نادیده گرفته می‌شود. کاو و همکاران در[۱۳]، به بررسی کاهش داده خط سیر با در نظرگرفتن اثرکاهش داده در پاسخ به پرسش‌های مکانی-زمانی پرداخته‌اند. فشرده‌سازی با استفاده از ساده‌سازی خط از یک تابع فاصله در تولید تقریب خط سیر استفاده می‌کند. در این روش کاهش داده، فاصله خط سیر اصلی از خط سیر تقریبی توسط یک پارامتر به نام حدآستانه خطا محدود می‌شود. با این حال حتی اگر خطای تقریب محدود باشد، خطای پاسخ به پرسش‌ها ممکن است نامحدود باشد. به عبارت دیگر، در جستجو روی خط سیرهای ساده شده، پاسخ‌ها ممکن است در مقایسه با نتایج به‌دست آمده از خط سیر اصلی، انحراف داشته باشد. بنابراین، برای تکنیک‌های کاهش داده خط سیر انحراف معرفی‌شده بوسیله ساده‌سازی خط باید مدیریت شود.

برای مدیریت خطای پاسخ به پرسش‌های مکانی-زمانی ناشی از ساده‌سازی خط، مفهومی به نام بی‌نقص بودن[۱] معرفی شده است. بی‌نقص بودن به این معنی است که اگر اختلاف پاسخ به پرسش‌های مکانی ـ زمانی روی خط سیر اصلی و خط سیر ساده‌سازی شده در یک محدوده مشخص باشد، آن تابع فاصله (که برای

[1] soundness

ساده‌سازی خط استفاده شده است) برای پرسش مربوطه بی‌نقص است. در[۱۳] نشان داده شده است که استفاده از ترکیب توابع فاصله E_t و E_u برای تمام پرسش‌های مکانی -زمانی روی خط سیر بی‌نقص است. در فشرده‌سازی خط سیر معنایی علاوه بر مفهوم بی‌نقص بودن، حفظ مناطق توقف و حرکت در خط سیر نیز مهم است. بر اساس مطالب بیان شده، روش پیشنهادی در بخش بعد معرفی می‌شود.

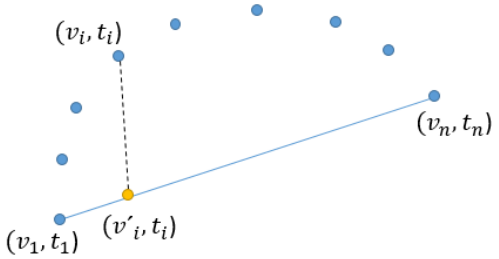
۴- روش پیشنهادی

در این تحقیق تمرکز روی حفظ نقاط توقف خط سیر معنایی در مسیر فشرده‌سازی داده‌ها است. روش پیشنهادی این پژوهش E_v_E_2 نامیده شده است. ایده اصلی این است که رفتار توقف- حرکت یک شئ بیشتر در مشتق آن، یعنی در سری‌های زمانی سرعت خط سیر، ظاهر می‌شود:

رابطه (۷)
$$E_v\big((x_i,y_i,t_i,v_i),(x_1,y_1,t_1,v_1),(x_n,y_n,t_n,v_n)\big)=\sqrt{(v_i-v_i')^2}$$

در رابطه ۷، E_v تابع فاصله بر اساس سرعت است. در این تابع فاصله، (x'_i,y'_i,t_i,v'_i) نقطه همزمان متناظر با (x_i,y_i,t_i,v_i) روی پاره خط واصل (x_1,y_1,t_1,v_1) و (x_n,y_n,t_n,v_n) است.

اما فشرده‌سازی صرفاً بر اساس سرعت کامل نیست. در این کار یک الگوریتم دو مرحله‌ای به نام E_v_E_2 ارائه شده است، که در مرحله اول از تابع فاصله E_v و در مرحله دوم از تابع E_2 برای اندازه‌گیری خطا استفاده می‌کند. در تابع خطا E_v هدف یافتن نقطه همزمان متناظر با (x_i,y_i,v_i) یعنی (x'_i,y'_i,v'_i) است و از فرض ثابت بودن شتاب بین ابتدا و انتهای خط سیر استفاده می‌کند. شکل ۶ درونیابی خطی بر اساس سرعت را در فضای سرعت -زمان را نشان می‌دهد. در این درونیابی، شیب خط واصل نقاط همان شتاب حرکت است. در درونیابی در تابع سرعت از فرض ثابت بودن شتاب حرکت استفاده شده است.



شکل۶- درونیابی خطی تابع فاصله Ev براساس شتاب ثابت

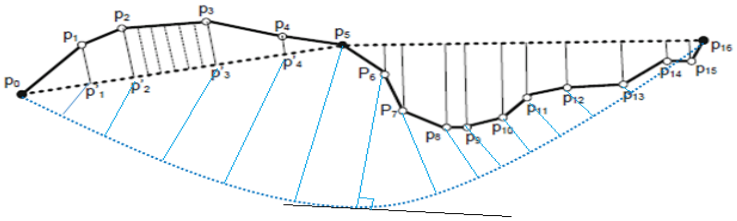
به این ترتیب درونیابی خطی نیز با همین فرض نقطه متناظر را تخمین می‌زند:

رابطه (۸)
$$v_i' = v_1 + a_{1n}.(t_i - t_1)$$

رابطه (۹)
$$a_{1n} = (v_n - v_1)/(t_n - t_1)$$

رابطه ۸، درونیابی خطی سرعت هر یک از نقاط را با فرض ثابت بودن شتاب در تابع E_v نشان می‌دهد و رابطه ۹ معادله محاسبه شتاب بر اساس نقاط ابتدایی و انتهایی خط سیر در هر تکرار برای رسیدن به مقدار بهینه است.

گام دوم در کاهش داده خط سیر با استفاده از الگوریتم E_v_E_2، محاسبه فاصله با استفاده از تابع E_2 است. تابع فاصله E_2 از فرض ثابت بودن سرعت در درونیابی استفاده می‌کند. همانطور که اشاره شد، تابع فاصله E_v از فرض شتاب ثابت استفاده می‌کند. بنابراین برای ترکیب این دو تابع، E_2 نیز باید بر اساس ثابت بودن سرعت بازنویسی گردد. به این ترتیب درونیابی در فضای مکانی به جای خط مستقیم تبدیل به سهمی می‌شود (شکل ۷).



شکل۷- تابع فاصله E2، مقایسه درونیابی با سرعت ثابت و شتاب ثابت

نکته دیگری که باید به آن توجه شود، موضوع هم مقیاس نبودن دو تابع فاصله E_2 و E_v به منظور مقایسه مقادیر است. ایده استفاده شده در این کار برای هم مقیاس کردن دو تابع فوق، تقسیم فواصل خطای تقریب هر یک از توابع به حد آستانه‌های تعریف شده است.

$$E_v\_ND = E_v/Th\_E_v \quad \text{رابطه (۱۰)}$$

$$E_2\_ND = E_2/Th\_E_2 \quad \text{رابطه (۱۱)}$$

در رابطه ۱۰، E_v_ND مقدار بدون واحد تابع فاصله E_v است و Th_E_v حد آستانه تعریف شده در تابع فاصله E_v برای شناسایی نقطه با بزرگترین فاصله در هر تقریب است. به همین ترتیب در رابطه ۱۱، E_2_ND مقدار بدون واحد E_2 است و Th_E_2 حد آستانه تعریف شده در تابع فاصله E_2 برای شناسایی نقطه با بزرگترین فاصله در هر تقریب است.

فرآیند فشرده‌سازی با استفاده از تابع $E_v\_E_2$ در شکل ۸ نشان داده شده است. همانگونه که مشاهده می‌شود، ابتدا داده های خط سیر ورودی که به‌صورت لیستی منظم از نقاط به شکل (x, y, t) وارد فرآیند محاسباتی می‌شود. سپس فاصله هر یک از نقاط خط سیر از اولین تقریب یعنی خط واصل نقاط ابتدا و انتهای خط سیر با تابع E_v محاسبه می‌شود. این محاسبات با فرض استفاده از تابع E_2 نیز برای نقاط خط سیر انجام می‌شود. به این ترتیب فاصله هریک از نقاط از سهمی ایجاد شده به واسطه ابتدا و انتهای نقاط، محاسبه می‌شود. با محاسبه فاصله نقاط با دو تابع E_v و E_2 بزرگترین فاصله برای مقایسه با حد آستانه تعیین شده و تصمیم‌گیری برای حفظ یا حذف نقطه مورد نظر، انجام می‌شود. در این تحقیق تعیین حدآستانه سرعت بر اساس بیشترین فاصله در تقریب اول محاسبه شده است. سپس الگوریتم با مقایسه رابطه بیشترین فاصله با حد آستانه‌های تعیین شده، نقطه مورد نظر را برای تقریب خط سیر ارزیابی می‌کند.

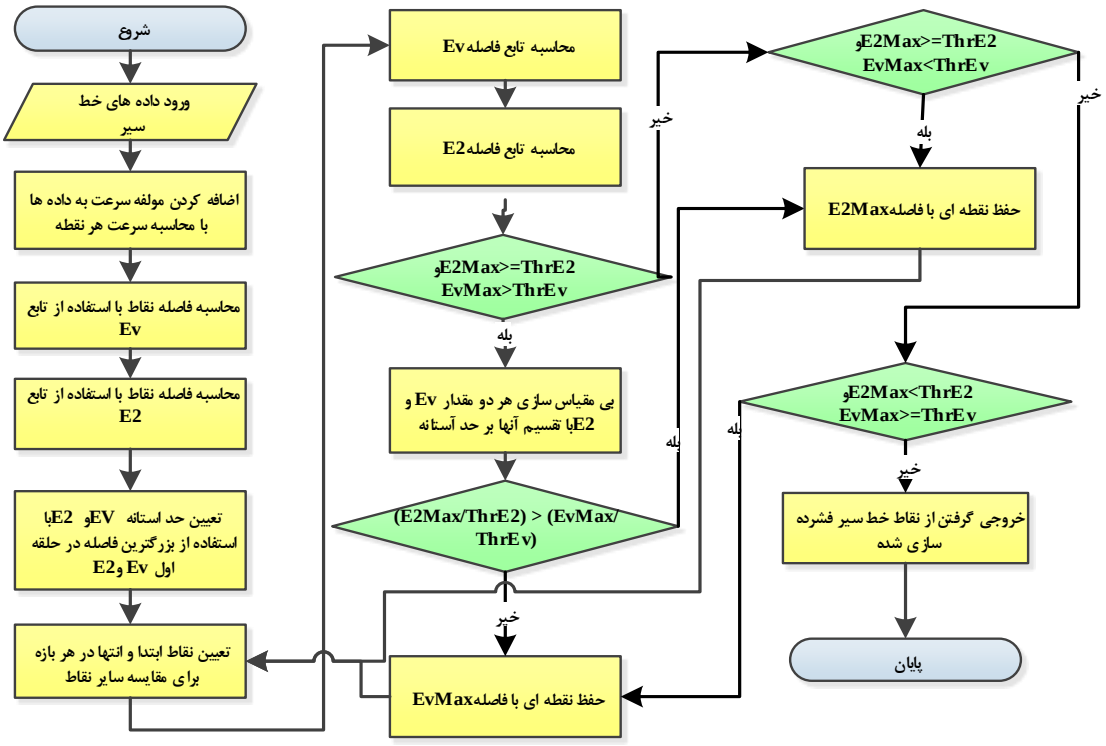
۵- پیاده‌سازی و ارزیابی نتایج

پیاده‌سازی تکنیک‌های فشرده‌سازی روی داده‌های خط سیر پاک سازی شده انجام می‌شود. برای این کار، از یک روش سرعت-مبنا برای شناسایی و حذف داده‌های پرت استفاده شده است. به این منظور، سرعت لحظه‌ای نقاط به صورت تقریبی با کمک دو نقطه قبل و بعد محاسبه شده است. سپس از قانون ۳-سیگما برای تعریف حد آستانه سرعت و شناسایی و حذف داده‌های پرت استفاده شده‌است.

برای پیاده سازی، الگوریتم فشرده سازی $E_v\_E_2$ از داده‌های خط سیر جئولایف [1] استفاده شده است[۱۴,۱۵,۱۶]این مجموعه داده شامل ۱۷۶۲۱ خط سیر است که توسط ۱۸۲ کاربر در طول ۵ سال (از آوریل ۲۰۰۷ تا آگوست ۲۰۱۲) جمع‌آوری شده است. داده‌های جمع‌آوری شده توسط دستگاه‌های مجهز به GPS در سیستم مختصات جغرافیاییWGS84 جمع‌آوری می‌شوند. از آنجائیکه در محاسبه تابع فاصله E_v (که گام اول در فشرده‌سازی به روش پیشنهادی است) خط سیر به صورت دنباله‌ای از سرعت نقاط نمایش می‌یابد، تبدیل مختصات از سیستم جغرافیایی به سیستم مختصات کارتزین ضروری است. بنابراین، ابتدا روی داده‌های مورد نظر تبدیل مختصات لازم

[1] GeoLife

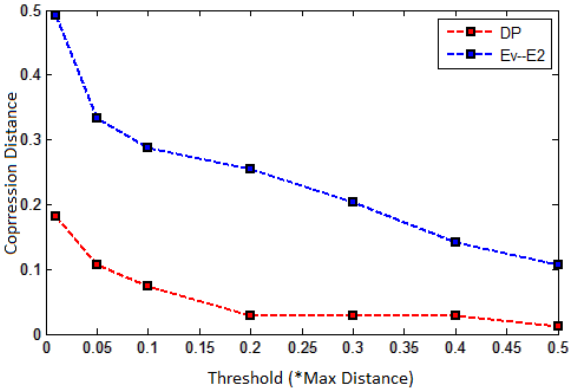
انجام گرفت تا داده‌ها از سیستم مختصات طول و عرض جغرافیایی به سیستم مختصات کارتزین تبدیل شوند.



شکل ۸- فرآیند فشرده‌سازی با استفاده از تابع Ev_E2

با پاکسازی داده‌ها از خطا و انجام تبدیلات لازم روی مختصات ورودی، الگوریتم پیشنهادی روی داده‌های خط سیر پیاده‌سازی شد. نتایج به دست آمده حاکی از موفقیت الگوریتم در کاهش تعداد نقاط، همزمان با حفظ مناطق توقف در خط سیر است. ارزیابی الگوریتم‌های فشرده‌سازی بر اساس پارامترهای مختلف می‌تواند انجام شود. یکی از معیارهای ارزیابی، نرخ فشرده‌سازی است. نرخ فشرده‌سازی برابر با نسبت تعداد نقاط حفظ شده حاصل از پیاده‌سازی الگوریتم فشرده‌سازی به تعداد کل نقاط در خط سیر است. الگوریتم فشرده‌سازی $E_v\_E_2$ با هدف حفظ مناطق توقف در خط سیر اجرا و پیاده‌سازی شده است. از این رو تعداد نقاط حفظ شده به مراتب از الگوریتم‌های فشرده‌سازی معمولی که بر اساس ساده‌سازی خطی هستند بیشتر است. شکل ۹ نرخ فشرده‌سازی روش پیشنهادی را در مقایسه با الگوریتم داگلاس-پوکر نشان می‌دهد. در این شکل، محور قائم بیانگر نرخ فشرده‌سازی و محور افقی تنظیمات مختلف مقادیر حد آستانه تعریف شده را بر اساس بیشترین فاصله نشان می‌دهد. همانطور که مشاهده می‌شود، نرخ فشرده‌سازی الگوریتم $E_v\_E_2$ نسبت به الگوریتم داگلاس-پوکر بیشتر است. با توجه به اینکه الگوریتم $E_v\_E_2$ یک الگوریتم ترکیبی است، نقاط حفظ شده بیشتر، اغلب نقاطی هستند که تابع E_v بر اساس فاصله در فضای سرعت-زمان آن‌ها را به‌عنوان نقاط مهم شناسایی و حفظ کرده‌است. بنابراین حفظ مناطق توقف در خط سیر به بهای افزایش نرخ فشرده‌سازی است. پارامتر دیگر برای ارزیابی دقت یک الگوریتم فشرده‌سازی پیچیدگی زمانی است. در بدترین حالت، زمان اجرای فشرده‌سازی با الگوریتم‌های ساده سازی خطی استاندارد (به عنوان مثال الگوریتم داگلاس-پوکر) O(n2) است. الگوریتم Ev_E2 نیز داری

همین پیچیدگی است. در واقع، اساساً الگوریتم Ev_E2 دو بار از الگوریتم‌های PLS استفاده کرده است.



شکل۹- مقایسه نرخ فشرده‌سازی روش داگلاس-پوکر (خط چین قرمز) و روش پیشنهادی (خط چین آبی)

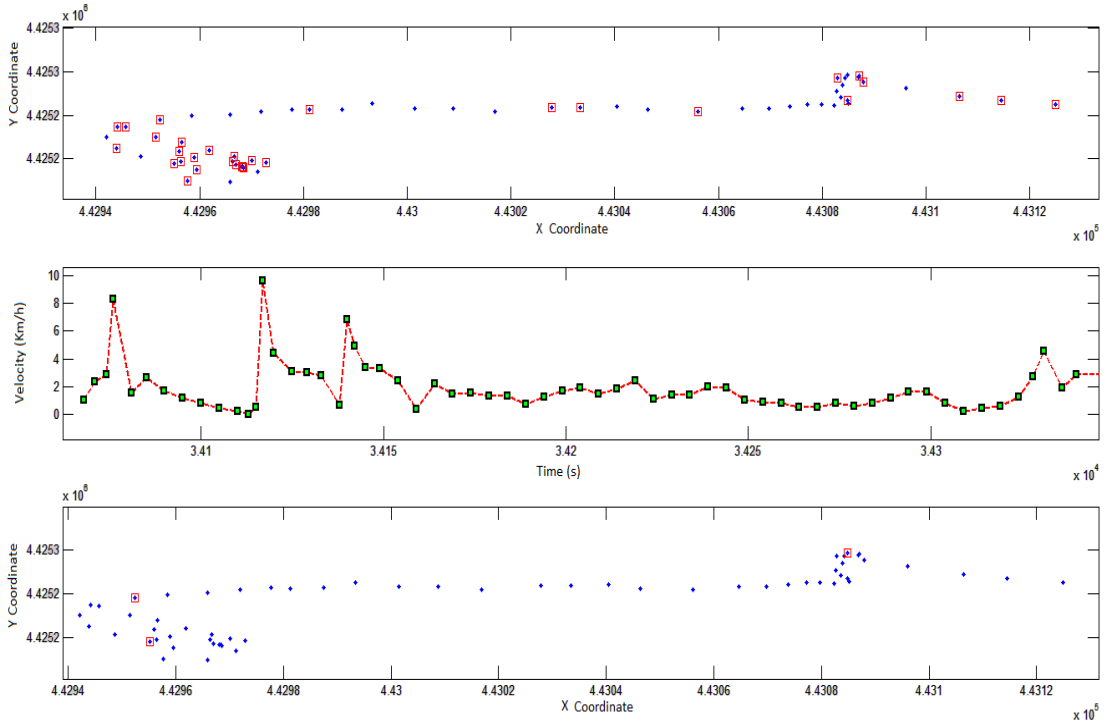
شکل ۱۰ نمونه‌ای از خط سیر فشرده‌سازی شده را نشان می‌دهد. در شکل ۱۰.(بالا) خط سیر فشرده‌سازی شده با الگوریتم پیشنهادی و در شکل ۱۰.(پایین) خط سیر فشرده شده با الگوریتم داگلاس-پوکر نشان داده شده است. در این شکل نقاط حفظ شده توسط دو الگوریتم با مربع قرمز نشان داده شده‌اند. همانطور که مشاهده می‌شود، مناطق توقف در خط سیر که اغلب به صورت تراکم مکانی نقاط دیده می‌شود، در الگوریتم پیشنهادی در مقایسه با الگوریتم داگلاس-پوکر به نحو مؤثرتری حفظ شده است. این نقاط مناطقی از خط سیر هستند که در آن شئ متحرک سرعت خود را کاهش داده است (شکل ۱۰.(وسط)). علاوه بر حفظ نقاط با سرعت پایین الگوریتم مناطقی از خط سیر که در آن شئ متحرک با سرعت بیشتر از میانگین حرکت کرده است را حفظ کرده که علت آن فاصله بیشتر این نقاط از تقریب خط سیر است.

۶-جمع‌بندی و نتیجه‌گیری

با تولید روزافزون داده‌های حرکت توسط دستگاه‌های تعیین موقعیت، حجم عظیمی از داده‌های مکانی زمانی خط سیر تولید شده است. در مطالعات مربوط به خط سیر مفهوم خط سیر معنایی در چند سال اخیر مورد توجه محققان این حوزه بوده است. در فرآیند غنی‌سازی معنایی خط سیرهای مکانی یک گام مهم در مدیریت داده ها، کاهش تعداد نقاط با حفظ دقت موردنیاز برای آنالیزهای مختلف است. با این حال اغلب تکنیک‌های کاهش داده بر اساس ماهیت مکانی زمانی خط سیر هستند و تمرکز روش‌های ارائه شده که اغلب بر اساس ساده‌سازی خطی هستند، حفظ نقاطی از خط سیر است که قادر به نمایش انحراف شئ متحرک از مسیر حرکت می باشد. در مفاهیم مربوط به خط سیر معنایی، مسیر شئ متحرک به صورت دنباله‌ای از مناطق توقف و حرکت نمایش می یابد. با این حال اغلب تکنیک‌های کاهش داده خط سیر قادر به حفظ مناطق و توقف نیستند.

در این مقاله یک روش کاهش داده خط سیر بر اساس سرعت نقاط پیشنهاد شده است که در آن تابع فاصله خط سیر مکانی-زمانی را به شکل یک سری زمانی از سرعت نقاط در نظر می گیرد. روش پیشنهادی که از ترکیب دو تابع فاصله E_2 و E_v برای تقریب استفاده کرده است، هدف حفظ نقاطی از خط سیر است که در فضای سرعت زمان بیشترین انحراف را از تقریب داشته باشد. از سوی دیگر نقاط با بیشترین انحراف در فضای مکانی-زمانی نیز به عنوان نقاط مهم در خط سیر حفظ می‌شوند.

روش پیشنهادی روی داده‌های خط سیر واقعی پیاده‌سازی شده است. نتایج به دست آمده از اجرای الگوریتم حاکی از فشرده‌سازی مؤثر داده‌های حرکت با حفظ مناطق و توقف در خط سیر است. با این حال این امر منجر به افزایش نرخ فشرده‌سازی و به دنبال آن افزایش پردازش‌های بعدی روی داده‌های خط سیر است. در فرآیند فشرده‌سازی خط سیر، نقاط با بیشینه و کمینه سرعت در فضای سرعت-زمان با تابع فاصله E_v و نقاط با بیشترین انحراف در فضای مکانی با تابع فاصله E_2 در خط سیر به عنوان نقاط مهم حفظ می‌شوند و از نقاط دیگر صرف نظر می‌شود.



شکل۱۰- مقایسه نتایج الگوریتم پیشنهادی (بالا)، نمودار سرعت نقاط (وسط) و نتایج حاصل از الگوریتم داگلاس پوکر (پایین)

الگوریتم‌های فشرده‌سازی با هدف کاهش حجم نقاط با حفظ دقت مورد نیاز برای تحلیل‌های مختلف روی داده‌ها اجرا می‌شوند. زمانی که هدف از تحلیل داده‌های خط سیر انجام آنالیزهای معنایی است، فرآیند فشرده‌سازی نقاط با رویکرد حفظ مناطق توقف در خط سیر انجام می‌شود. به این ترتیب نتایج حاصل از این تحقیق به صورت مستقل از نوع داده‌های جمع‌آوری شده می‌تواند در تحلیل خطوط سیر معنایی مورداستفاده قرار گیرد.

مراجع

[1] Alvares, L. O., Oliveira, G., Heuser, C. A., and Bogorny, V., "A Framework for Trajectory Data Preprocessing for Data Mining," in Conf. on Software Engineering and Knowledge Engineering, 2009.

[2] Pelekis, N., Theodoridis, Y., Janssens, D., "On the Management and Analysis of Our LifeSteps," ACM SIGKDD Explorations Newsletter, vol. 15, no. 1, pp. 23-32, 2013.

[3] Zheng, Y., , Zhou, X., Computing with Spatial Trajectories, New York: Springer-Verlag New York, 2011.

[4] Douglas, D., Peucker, T., "Algorithms for the Reduction of the Number of Points Required to Represent a Line or its Caricature," Cartographica, vol. 10, no. 2, pp. 112-122, 1973.

[5] Maratnia, N., de By, R., "patio-Temporal Compression Techniques for Moving Point Objects," in International Conference on Extending Database Technology (EDBT), 2004.

[6] R. Bellman, "On the Approximation of Curves by Line Segments Using Dynamic Programming," Communications of the ACM, vol. 4, no. 6, 1961.

[7] J. Vitter, "Random sampling with a reservoir," ACM Transactions on Mathematical Software(TOMS), vol. 11, no. 1, 1985.

[8] Keogh, E., Chu, S., Hart, D., Pazzani, M., "An On-Line Algorithm for Segmenting Time Series," in International Conference on Data Mining (ICDM), 2001.

[9] Potamias, M., Patroumpas, K., Sellis, T., "ampling Trajectory Streams with Spatio-Temporal Criteria," in In: International Conference on Scientific and Statistical Database Management (SSDBM), 2006.

[10] K. Patroumpas, "Online tracking and summarization over streaming maritime trajectories," in Workshop on Moving Objects at Sea, 2013.

[11] W. Ting, "SOMETIMES TOO BIG: COMPRESSING TRAJECTORY DATA," in PACIS 2014 Proceedings, 2014.

[12] de Vries, G. K. D. and ,Someren, M., "Machine learning for vessel trajectories using compression, alignments and domain knowledge," ELSEVIER, vol. 39, no. 18, p. 13426–13439, 2012.

[13] Cao, H., Wolfson, O. and Trajcevski, G., "Spatio-temporal data reduction with deterministic error bounds," The VLDB Journal — The International Journal on Very Large Data Bases , vol. 15, no. 3, pp. 211-228, 2006.

[14] Zheng, Y., Zhang, L., Xie, X., & Ma, W. Y., "Mining interesting locations and travel sequences from GPS trajectories," in In Proceedings of the 18th international conference on World wide, Madrid Spain, 2009.

[15] Zheng, Y., Li, Q., Chen, Y., Xie, X., & Ma, W. Y., "Understanding mobility based on GPS data," in In Proceedings of the 10th international conference on Ubiquitous computing, Seoul, Korea, 2008.

[16] Zheng, Y., Xie, X., & Ma, W. Y., "GeoLife: A Collaborative Social Networking Service among User, location and trajectory," in IEEE Data Engineering Bulletin, 2010.

K. N. TOOSI UNIVERSITY OF TECHNOLOGY
FACULTY OF GEODESY AND GEOMATICS ENGINEERING

Journal of Geospatial Information Technology
Vol.3, No.4, Winter 2016

An algorithm for compression of a spatio-temporal trajectory preserving its semantic nature

Somaie Aghel Shahneshin [1], Simin Sadat Mirvahabi [1], Rahim Ali Abbaspor *[2]

1- MSc student, School of Surveying and Geospatial Information Engineering, College of Engineering, University of Tehran
2- Assistant professor, School of Surveying and Geospatial Information Engineering, College of Engineering, University of Tehran

Abstract

A common way to store information of spatio-temporal moving objects is to display the path of the objects as the form of a three-dimensional trajectory using the geographic location and time. In recent years, extensive research has been done on the trajectories. These studies have focused mainly on geometric aspects of trajectories. However, semantic trajectory is a relatively new concept that has been developed with the purpose of effective semantic analysis on captured data. In semantic trajectory, which is a secondary display of geometric trajectory, the movement of object is described as series of stop-and-move. Production of semantic trajectory from the collected raw data is a process with several steps. Due to the huge amount of data, one of the important processes is reducing the number of points of trajectory with maintaining the required accuracy by using compression techniques. However, data reduction techniques commonly are based on linear simplification and are not able to protect stop and move of trajectories. In this paper, a data reduction technique is presented which is based on combination of two distance functions for approximation of semantic trajectory. The first distance function has used speed of points to calculate the approximation error of trajectories. The second function is based on the development of well-known Douglas-Peuker algorithm, which assumes constant acceleration to calculate the approximation error. The proposed algorithm is implemented on real trajectory data and the results show improved performance compared with other algorithms in preservation of the stop and move of trajectories.

Key words: Semantic Trajectory, compression, Stop-Move Model, E_v-E_2

Correspondence Address GIS Group, School of Surveying and Geospatial Information Engineering, College of Engineering, University of Tehran, Tehran, Iran. **Tel: +98 21 88008841.**
Email: abaspour@ut.ac.ir