

خلاصه‌سازی جریان داده خط‌سیر مبتنی بر گراف

امیرحسین شکوهی^۱، عین‌اله خنجری میانه^۲.

۱- دانشجوی دکترا، آزمایشگاه مهندسی داده، دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران

۲- استادیار، آزمایشگاه مهندسی داده، دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران

چکیده

داده‌های مکانی زمانی خودروهای در حال حرکت که به صورت برخط و جریانی ارسال می‌شود دارای سرعت و نرخ بالایی بوده و حجم زیادی را اشغال می‌کنند. بنابراین محققان به دنبال روش‌هایی هستند تا بتوان بدون از دست دادن روند داده و خط‌سیر اشیا متحرک، آن‌ها را خلاصه‌سازی کرده و برای تحلیل جریان داده خط‌سیر مورد استفاده قرار دهند. روش‌های مختلفی توسط محققان حوزه خط‌سیر برای خلاصه‌سازی داده‌های مکانی زمانی اشیا متحرک پیشنهاد گردیده است. در این پژوهش راهکاری ارائه خواهد شد تا داده‌های خط‌سیر که به صورت جریانی و با نرخ بالایی تولید می‌شوند؛ بر یک ساختار گرافی منطبق شده به گونه‌ای که یال‌های گراف نشان‌دهنده جریان پویای اشیا متحرک و مبتنی بر زمان بوده و حجم ترافیک بین نواحی مکانی مختلف به صورت یال‌های وزن‌دار جهت‌دار نمایش داده شوند. رأس‌های گراف یک مجموعه ثابت بوده اما یال‌ها براساس حجم و جریان ترافیک می‌توانند تغییر کنند. این مقاله جریان داده خط‌سیر مربوط به اشیا متحرک را به صورت پویا و مبتنی بر گراف خلاصه‌سازی و توصیف کرده تا بتوان از آن در تحلیل اطلاعات رفتاری و الگوهای حرکتی اشیا متحرک مورد استفاده قرار داد.

کلمات کلیدی: شی متحرک، خط‌سیر، گراف پویا، خلاصه‌سازی جریان داده

۱. مقدمه

خلاصه‌سازی داده‌ها یکی از وظایف و بخش‌های مهم تحلیل داده‌هاست که می‌تواند به صورت فشرده‌سازی داده‌ها تعریف شود. خلاصه‌سازی، جزئیات غیرمرتبط را از داده‌ها حذف کرده که فقط اطلاعات مهم باقی بماند. در برخی از منابع، خلاصه‌سازی داده‌ها را به صورت توانایی مدیریت موثر حجم بزرگی از اطلاعات پیچیده بیان کرده‌اند [1]. این مقاله به دنبال خلاصه‌سازی داده‌های حرکتی است. حرکت اشیا متحرک باعث تولید حجم زیادی از داده‌های مکانی-زمانی می‌شوند. داده‌های حرکتی اشیا متحرک، خط‌سیر مکانی نامیده می‌شوند که دنباله‌ای از مختصات‌های مکانی بوده و در بازه‌های زمانی پشت سرهم تولید می‌شوند.

پیشرفت‌های فناوری در موقعیت‌یابی و تجهیزات ارتباطی از عوامل تجمع حجم عظیمی از داده‌های خط‌سیرهای مکانی است. خط‌سیرهای مکانی را به دلیل حجم بالای آن‌ها نمی‌توان به راحتی مدیریت کرد [1]. یکی از مشکلات، استخراج



اطلاعات وابسته به زمان برای هر حرکت مشخص و استفاده از این اطلاعات به صورت لحظه‌ای و پویاست. اگر بتوان برخی از اطلاعات را استخراج و فشرده‌سازی کرد - مانند خط‌سیر خلاصه شده- آنگاه مجموعه داده‌ها بسیار کوچک‌تر شده و دستیابی به اطلاعات مرتبط آسان‌تر خواهد شد.

این تحقیق روشی را برای خلاصه‌سازی داده‌های خط‌سیر مبتنی بر گراف پیشنهاد خواهد داد. نمایش گرافی یک خط سیر، روشی قدرتمند و سودمند برای توضیح روابط بین اجزای داده‌ها است. مطالعه و بررسی این گراف‌ها اطلاعات مفیدی درباره روابط و ساختار داده‌های مکانی-زمانی ارائه می‌کند. با توجه به اینکه اطلاعات خط‌سیر وابسته به زمان می‌باشد، از گراف‌های وابسته به زمان استفاده خواهد شد که این گراف‌ها به صورت پویا و دینامیک در طول زمان تغییر خواهند کرد. در این مدل‌سازی، رأس‌های گراف، ناحیه‌های مکانی شبکه حمل و نقل بوده و جریان ترافیکی بین این ناحیه‌های مکانی، یال‌های گراف می‌باشند. این یال‌ها جهت‌دار بوده و با توجه به اینکه گراف وابسته به زمان می‌باشد در بازه‌های زمانی مختلف یال‌ها دارای وزن‌های مختلفی نیز خواهند بود. بنابراین برخلاف روش‌های سنتی که برای توصیف خط‌سیر از گراف‌های ایستا استفاده می‌کردند، در رویکرد پیشنهادی با توجه به ماهیت داده‌های خط‌سیر که به صورت متغیر با زمان و جریانی تولید می‌شوند، گراف‌های پویا، ایجاد می‌شود که با این کار نقاط تغییر در گراف در بازه‌های زمانی مختلف قابل مشاهده خواهد بود. به این ترتیب در پنجره‌های زمانی مختلف، یال‌هایی با وزن‌ها و جهت‌های مختلف به گراف اضافه یا حذف خواهند شد. گراف‌های ایجاد شده مرتبط با هر پنجره زمانی، به ترتیب زمانی قرار گرفته و دنباله‌ای از گراف‌ها را تولید می‌کنند. با این روش برای هر بازه زمانی یک گراف تولید شده و کل داده‌های خط‌سیر مربوط به آن بازه زمانی خلاصه شده و به صورت یک گراف نمایش داده می‌شود. بنابراین نه تنها تحلیل و بررسی و توصیف داده‌های خط‌سیر، ساده‌تر خواهد بود بلکه الگوی رفتارهای ترافیکی و توزیع مکانی ترافیک در بازه‌های زمانی مشابه در روزهای مختلف استخراج می‌گردد.

۲. ناحیه‌بندی مکانی

به منظور تبدیل داده‌های خط‌سیر به گراف زمانی، اولین قدم مشخص کردن ناحیه‌های مکانی است که خط‌سیر از این نواحی عبور می‌کند. خط‌سیر یک شی متحرک دنباله‌ای از نقاط $P(x,y,t)$ به ترتیب زمان است که x,y مختصات مکانی شی در زمان t است. یک روش ساده برای ناحیه‌بندی مکانی که در کارهای پیشین استفاده کرده‌اند، استفاده از حومه نقاط یا همان مرزهای کدپستی محلات است که خط‌سیر از آن مناطق عبور می‌کند. در این پژوهش روشی که برای ناحیه‌بندی مکانی استفاده خواهد شد براساس روش ورونوی [3] یا دیاگرام ورونوی، خواهد بود. با استفاده از این روش، به هر مجموعه از نقاط روی خط‌سیر، ناحیه‌ای اختصاص داده می‌شود. در حقیقت با استفاده از روش ورونوی مرزهای هر ناحیه مکانی تعیین می‌شود. پس از انتخاب مجموعه‌ای از نقاط خط‌سیر به عنوان نقاط ایجادکننده نواحی با استفاده از روش ورونوی، خط‌سیر به بخش‌هایی تقسیم می‌شود به طوری که براساس روش ورونوی هر نقطه بر روی خط‌سیر به نقطه ایجادکننده ناحیه خود نزدیک‌تر از ایجادکننده سایر نواحی خواهد بود. پس از انجام این تقسیم‌بندی، در بخش آتی روش بهبود این تقسیم‌بندی بیان خواهد شد.

نقاط ایجادکننده را به صورت از پیش تعریف شده در نظر گرفته و مرز هر ناحیه مکانی براساس دیاگرام ورونوی تعیین می‌شود. همانطور که بیان گردید؛ نقاط موجود در هر ناحیه، کمترین فاصله را به نقطه ایجادکننده همان ناحیه نسبت به نقاط ایجادکننده سایر نواحی خواهند داشت و نقاطی که در مرز دو ناحیه قرار دارند، فاصله یکسانی از نقاط ایجادکننده دو ناحیه دارند. در مرحله تکمیلی، به جهت بهبود ناحیه‌بندی، نواحی ایجاد شده یا ادغام شده یا به نواحی جدیدتر تجزیه میشوند تا زمانی که نواحی همگن ایجاد شود. برای ایجاد نواحی همگن از یک روش بخش‌بندی خط‌سیرها، معروف به روش GRASP استفاده خواهد شد [2]. در این تحقیق از روش مذکور به صورت اصلاح شده و بهبود یافته استفاده خواهد شد.

روش اصلاح شده GRASP به این گونه است که نقاط نماینده هر ناحیه مکانی ایجاد شده در مرحله قبل توسط روش ورونوی، به عنوان نقاط نماینده مطرح شده در روش GRASP با عنوان LandMark در نظر گرفته می‌شود. برای برقراری همگنی حداکثری بین نقاط یک ناحیه، از اصل MDL استفاده می‌شود. اصل MDL یا همان تابع هزینه، میزان شباهت نقاط نماینده هر ناحیه با سایر نقاط آن ناحیه و همچنین میزان شباهت نمایندگان هر ناحیه مکانی را محاسبه می‌کند [4]. به این ترتیب با حذف، اضافه و یا جابجا شدن نقاطی که تغییرات مهم در آن صورت گرفته است، و همچنین جابجایی نمایندگان هر ناحیه، نواحی همگن ایجاد شده و بهترین نواحی برای تولید گراف زمان‌دار متناظر با خط سیر فراهم می‌شوند. برای بررسی میزان شباهت بین نقاط، معیارهای مختلفی را می‌توان در نظر گرفت. در این مقاله معیار شباهت نقاط موجود بر روی خط سیر قرار گرفته در یک ناحیه مکانی، فاصله اقلیدسی بین آن نقاط است. فرض کنید $T = \{s_1, s_2, \dots, s_T\}$ یک بخشبندی از خط سیر T باشد، میزان شباهت بین دو نقطه P_i و P_j از T عبارت است از:

$$Similarity(P_i, P_j) = \sqrt{\sum_{i=1}^N (P_i - P_j)^2}$$

و برای نماینده lm_i از ناحیه i ام که به صورت $\langle P_u, \dots, P_v \rangle$ می‌باشد، میزان شباهت نماینده ناحیه از سایر نقاط ناحیه متناظر با آن عبارت است از:

$$\sum_{k=u}^v Similarity(lm_i, P_k)$$

با توجه به روابط فوق، تابع هزینه مربوط به اصل MDL به صورت زیر بیان می‌گردد:

$$cost_function(\alpha_T, \beta_T) = f(\alpha_T) + g(\beta_T)$$

که در رابطه فوق تابع f و g به صورت زیر تعریف شده است:

$$f(\alpha_T) = \log_2\left(\sum_{j=1}^{T-1} Similarity(lm_j, lm_{j+1})\right)$$

$$g(\beta_T) = \log_2\left(\sum_{i=1}^T Similarity(lm_i, s_i)\right)$$

که تابع $f(\alpha_T)$ در حقیقت میزان شباهت بین نقاط نماینده نواحی مختلف یا همان LandMarkها را مشخص کرده که در اصل MDL بایستی شباهت این نقاط بسیار کم بوده و تابع $g(\beta_T)$ که شباهت نقاط داخل یک ناحیه مکانی را می‌سنجد بایستی حداکثر بوده و این نقاط بسیار شبیه هم باشند.

۳. ایجاد گراف پویا

در بخش قبل به روش تقسیم خط‌سیر به ناحیه‌های مکانی اشاره شد. در این بخش با استفاده از تقسیم خط‌سیر، گراف متناظر با خط‌سیر ایجاد می‌گردد. برای این منظور، کل زمان حرکت شی متحرک به T بازه زمانی با طول‌های ثابت Δt تقسیم می‌شود. هر بازه زمانی یک پنجره زمانی نام دارد. در این بخش برای هر پنجره زمانی یک گراف خط‌سیر به صورت

$$G_t = (V_t, E_t \subseteq (V \times V))$$

که $t = 1, 2, \dots, T$ متناظر می‌گردد و V_t مجموعه رئوس گراف، نماینده‌های نواحی مکانی در پنجره زمانی t ام هستند که در بخش قبل نحوه استخراج آن‌ها بیان گردید. E_t مجموعه یال‌های وزن‌دار و جهت‌دار در پنجره زمانی t ام هستند که بین هر دو رأس متوالی یالی جهت‌دار قرار می‌گیرد و وزن هر یال به صورت یک بردار دوتایی (\vec{v}, \vec{v}') است که \vec{v} متوسط سرعت و \vec{v}' متوسط انحنا مسیر بین دو رأس متوالی می‌باشد.

با ساخت گراف متناظر در هر پنجره زمانی، دنباله‌ای وابسته به زمان از گراف‌ها تولید می‌شود. این گراف‌ها توصیف ساده‌ای از نحوه حرکت شی متحرک در هر پنجره زمانی ارائه می‌کنند که با مقایسه گراف‌ها در پنجره‌های زمانی متوالی، می‌توان میزان تغییر در حرکت شی متحرک را بررسی کرد.

یکی از مزایای استفاده از نمایش مبتنی بر گراف این است که می‌توان دیدگاه‌های خلاصه بیشتری را از داده‌های اصلی به وسیله روش‌های خلاصه‌سازی و فشرده‌سازی گراف‌ها به دست آورد. پس از ایجاد گراف‌های متعدد در بازه‌های زمانی متوالی، بایستی تعداد گراف‌های ساخته شده را کاهش داد به گونه‌ای که بهینه‌سازی در عملیات خلاصه‌سازی صورت گیرد. برای این منظور برای هر پنجره زمانی، یک تابع تغییرات نسبت داده می‌شود که متوسط تغییرات در سرعت و انحنا را نشان می‌دهد. با محاسبه این تابع در هر پنجره زمانی می‌توان نقاطی را که تابع تغییرات به مقدار چشمگیری افزایش یا کاهش می‌یابد مشخص کرد. اصطلاحاً این نقاط را نقاط تغییر^۱ می‌نامند. در فاصله بین هر دو نقطه تغییر متوالی، چندین گراف وجود دارد که شباهت زیادی به هم دارند. لذا این گراف‌ها را با هم ادغام کرده و کلیه این گراف‌های مشابه از لحاظ تابع تغییرات، در یک گراف ادغام می‌شوند. به این ترتیب تعداد گراف‌ها به طور محسوسی کاهش یافته و نهایتاً توصیف ساده‌تری از خط‌سیر حاصل می‌گردد.

۴. نتیجه‌گیری

در این پژوهش یک چارچوب مبتنی بر گراف برای توصیف تحلیل خط‌سیر پیشنهاد گردید، که یک الگوی مکانی-زمانی برای خط‌سیر ارائه می‌دهد. به منظور درک و تحلیل بهتر خط‌سیر، دنباله‌ای از گراف‌های زمان‌دار متناظر با خط‌سیر در نظر گرفته شد.

هدف چارچوب پیشنهادی، به کارگیری روش‌های داده‌کاوی و گراف‌کاوی به همراه یک مجموعه غنی از داده‌های واقعی خط‌سیر برای فراهم کردن روشی جدید به منظور فهم، تحلیل و بصری‌سازی جریان ترافیکی در خط‌سیر می‌باشد. در این مقاله رویکردی برای بخش‌بندی خط‌سیر به ناحیه‌های مکانی ارائه گردید و بر روی چندین روش برای ناحیه‌بندی خط‌سیر در هر پنجره زمانی بحث شد و سرانجام با تعیین و به کارگیری نقاط تغییر مهم در خط‌سیر، دنباله‌ای از

¹ Change point

گراف‌ها، به یک مجموعه از گراف‌ها تبدیل می‌شود. این نمایش گرافی توصیف کامل و پر از مفهوم خلاصه شده‌ای از الگوی وابسته به زمان داده‌های اصلی خط سیر ارایه می‌دهد و اجازه می‌دهد تا یک بیان قوی از خط سیر برای کاربردهای مختلف مانند تشخیص بی‌نظمی و پیش‌بینی ازدحام ارایه گردد.

۱۲. مراجع

- 1.M.L. Damiani, F. Hachem, Segmentation techniques for the summarization of individual mobility data, WIREs Data Min. Knowl. Discov. (2017), <https://doi.org/10.1002/widm.1214>.
2. A. Soares J´unior, B. N. Moreno, V. C. Times, S. Matwin, and L. A. F.Cabral. GRASP-UTS: an algorithm for unsupervised trajectory segmentation. *Int. J. of Geographical Information Science*, 29(1):46–68, 2015.
3. Aurenhammer, F., 1991. Voronoi Diagrams—a Survey of a Fundamental Geometric Data Structure. *ACM Comput. Surv.* 23, 345–405. doi:10.1145/116873.116880.
4. P. D. Grunwald, I. J. Myung, and M. Pitt. *Advances in Minimum Description Length*. MIT Press, 2005.
5. Kim, J., K. Zheng, S. Ahn, M. Papamanolis, and P. Chao. Graph-Based Analysis of City-Wide Traffic Dynamics Using Time-Evolving Graphs of Trajectory Data. Proc., 38th Australasian Transport Research Forum (ATRF), Melbourne, Victoria, Australia, 2016.
6. M. Buchin, A. Driemel, M. Van Kreveld, and V. Sacristan. Segmenting trajectories: A framework and algorithms using spatiotemporal criteria. *Journal of Spatial Information Science*, 3(3):33–63, 2011.
7. V. Bogorny, C. Renso, A. R. de Aquino, F. de Lucca Siqueira, and L. O. Alvares. CONSTAnT a conceptual data model for semantic trajectories of moving objects. *Trans. in GIS*, 18(1):66–68, 2014.