4th International Conference on Web Research
April 25-26, 2018; Tehran, IRAN
ICWR2018
کنفرانس بین المللی وب پژوهی
چهارمین
۵ و ۶ اردیبهشت ۱۳۹۷؛ تهران، ایران

بررسی روش ها و کاربردهای تحلیل داده‌های مکانی

مریم ضابط[۱]، میلاد عدلی[۲]، بهشید بهکمال[۳]

[۱] کارشناس ارشد نرم افزار، دانشگاه فردوسی مشهد، ایران
[۲] دانشجوی کارشناسی ارشد نرم افزار، دانشگاه فردوسی مشهد، ایران
[۳] استادیار گروه مهندسی کامپیوتر دانشگاه فردوسی مشهد، ایران

چکیده

داده‌های مکانی، به داده‌هایی گفته می‌شود که بیان‌کننده موقعیت جغرافیایی یک موجودیت (طبیعی یا مصنوعی) باشند. مهمترین ویژگی های داده های مکانی که باعث پیچیدگی بسیار بالای کاوش داده های مکانی می شوند حجم بسیار زیاد، روابط و قوانین پیچیده حاکم بر این داده هاست که از ترکیب اطلاعات مکانی، توصیفی و زمانی حاصل می شود. در این پژوهش ابتدا روش‌های تحلیل و کاوش داده‌های مکانی طبقه بندی می‌شود. سپس راهکارهای ارائه شده در مقالات گذشته براساس طبقه بندی ارائه شده مورد بررسی قرار می گیرد و در پایان براساس ویژگی ها و نقاط قوت راهکارهای ارائه شده، ارزیابی مقایسه‌ای از کاربردهای مورد بررسی ارائه خواهد شد.

کلمات کلیدی

داده‌های مکانی، تحلیل داده، طبقه‌بندی

۱- مقدمه

داده‌کاوی یافتن الگوهای غیر بدیهی، ضمنی، مفید و نهفته در داده‌هاست. این الگوها کمک می‌کنند تا از نتایج به‌دست‌آمده برای تصمیم‌گیری در راهبری سازمان‌ها استفاده کرد. با توجه به حجم بالای داده‌ها امکان ایجاد نتایج از روش‌های غیرمکانیزه عملا غیرممکن و یا نیازمند تجربه بالای تصمیم‌گیرنده است. با توجه به این محدودیت در زمینه حجم داده‌ها، استفاده از روش‌ها و الگوریتم‌های داده‌کاوی مهم‌ترین عامل موفقیت در رقابت های سازمانی محسوب خواهد شد.

عمده این الگوریتم‌ها در دسته هندسه محاسباتی قرار می گیرند [17]. این دسته از الگوریتم‌ها به بررسی مسائلی می پردازند که در آن داده‌های هندسی وجود دارند. یکی از مهم‌ترین و متداول‌ترین انواع این داده‌ها، داده‌های مکانی (مکانی) هستند. داده‌های مکانی نه تنها ارتباط ویژه‌ای با الگوریتم‌های هندسه محاسباتی از حیث هندسی‌بودن دارند بلکه از نظر ویژگی‌های داده‌ای دیگر که مورد نظر روش های داده‌کاوی هستند نیز کاملا در حوزه روش‌های داده‌کاوی با حجم بالا قرار می گیرند.

به عنوان نمونه در یک شهر با یک میلیون نفر جمعیت، با احتساب تمام عوارض مکانی مانند املاک، شبکه‌های تأمین زیرساخت (آب، برق، گاز، تلفن و ...)، راه ها، دارایی‌های شهری (درخت ها، نیمکت‌های شهری، تابلوها و ...)، بیش از ده میلیون رکورد اطلاعاتی وجود خواهد داشت. در صورتی که اطلاعات رویدادهای مختلف مانند موقعیت مکانی خودرو ها در طول روز که به فاصله هر ۱۵ ثانیه برداشت شده اند نیز به این مجموعه اضافه شود، روزانه بیش از ده‌ها میلیون رکورد اطلاعاتی جدید تولید می شود [18,19].

همچنین این داده‌ها دارای روابط بسیار پیچیده‌ای می باشند. این روابط شامل قوانین شناخته‌نشده و محدودیت‌های هندسی برروی داده‌ها می باشند. با توجه به محدودیت انسان‌ها در طی مسیر و همچنین زمان، چنانچه رویدادها به اطلاعات مکانی متصل شوند، می‌توانند روابط و داده‌های پنهان بسیار زیادی را نمایان سازند. همچنین اعمال محدودیت‌های هندسی و مکانی برروی رکوردهای اطلاعاتی و همچنین رعایت این قوانین و محدودیت‌ها خود می تواند پیچیدگی‌های زیادی به مسئله اضافه کند.

در این تحقیق به منظور بررسی ابعاد مختلف داده های مکانی، ابتدا با مطالعه کارهای گذشته چالش های موجود شناسایی شده و راهکارهای ارائه شده در مقالات مختلف مورد بررسی قرار گرفت. سپس کاربردهای کاوش داده های مکانی بر اساس دسته بندی ارائه شده در [24] طبقه بندی شد. همچنین

Archive of SID
www.SID.ir

ویژگی هایی که در هر یک مقالات به عنوان نقاط قوت روش پیشنهادی مطرح شده استخراج شد تا براساس آن ارزیابی مقایسه ای کارهای موجود انجام شود. در ادامه مقاله، در بخش دوم روش‌های تحلیل و کاوش داده‌های مکانی طبقه بندی می شود. سپس در کاربردهای کاوش داده های مکانی براساس طبقه بندی ارائه شده مورد بررسی قرار می گیرد. در پایان نیز ارزیابی مقایسه‌ای از کاربردهای مورد بررسی ارائه خواهد شد.

۲- روش‌های کاوش داده های مکانی

داده‌های مکانی موجود در پایگاه‌های داده، به خصوص در سیستم‌های اطلاعاتی جغرافیایی[1] حجم داده‌ای بسیار زیادی تولید می‌کنند. این حجم بسیار زیاد داده‌ها شامل داده‌های مفید ، غیرمفید و الگوهای پنهان می باشد که می‌توانند با استفاده از روش‌های داده کاوی، اصلاح و قابل استفاده شوند. به عنوان نمونه در خصوص داده های مکانی مسیر حرکت یک خودرو می توان به مسیر حرکت به عنوان داده های مفید، خطاهای موجود در برداشت موقعیت مکانی و همچنین داده های تکراری ناشی از توقف به عنوان داده های غیرمفید و زاید و الگوهای حرکتی فرد در طول روزهای مختلف به عنوان الگوهای پنهان اشاره نمود.

به طور کلی، کشف دانش از پایگاه‌داده‌های مکانی به معنای استخراج دانش ضمنی، ارتباطات مکانی و کشف الگوهای جالبی است که به طور صریح در پایگاه‌های داده نمایش داده نمی‌شوند. این تکنیک‌ها نقش مهمی در فهم داده‌های مکانی و دستیابی به روابط درونی بین داده‌های مکانی و غیرمکانی دارند. علاوه بر این، می توان از روابط کشف‌شده برای نمایش داده‌ها به صورت مختصر، بازسازی پایگاه‌های داده مکانی و همچنین بالا بردن کارایی استفاده کرد. کاوش داده های مکانی[2] کاربرد زیادی در زمینه‌های مختلف مانند سیستم‌های اطلاعاتی جغرافیایی، اکتشاف پایگاه‌داده تصویری، تصویربرداری پزشکی و غیره دارد.

کاوش داده های مکانی یک زمینه تحقیق در حال رشد است که هنوز در مراحل اولیه است. در طول دهه گذشته با توجه به کاربرد گسترده تکنولوژی سیستم مکان یابی جهانی[3]، نگاشت و اشتراک داده‌های مکانی مبتنی بر وب، سنجش از راه دور با وضوح بالا و همچنین خدمات مبتنی بر موقعیت مکانی، دامنه‌های تحقیقاتی بیشتری به داده‌های جغرافیایی برای تجزیه و تحلیل‌های اجتماعی و کاربردهای تجاری، دست پیدا کرده‌اند. بنابراین پیش‌بینی می شود که استفاده‌های جدیدتری از داده های مکانی خواهد شد و روش‌های کاوش داده های مکانی جدیدی توسعه خواهند یافت. کاوش داده های مکانی شامل وظایف مختلفی است که برای هر منظور چندین روش موجود است که می تواند محاسباتی، آماری، بصری و یا ترکیبی از آن ها باشد. روشهای کاوش داده‌های مکانی را می توان براساس نوع وظیفه داده‌کاوی به چهار دسته زیر طبقه‌بندی کرد:

۲-۱-۱- طبقه بندی و پیش بینی مکان

روش‌های طبقه‌بندی مکانی، گسترش‌یافته‌ی روش‌های طبقه‌بندی عمومی هستند تا علاوه بر ویژگی‌های شیء ، ویژگی‌های اشیاء همسایه و روابط مکانی آن‌ها را نیز طبقه‌بندی کنند [1]. یک روش بصری طبقه‌بندی در [2] معرفی شده است که در آن درخت تصمیم مشتق‌شده با الگوریتم سنتی C4.5 با تجسم نقشه ای ترکیب شده تا بتواند الگوهای مکانی قوانین طبقه‌بندی را آشکار کند.در [3] استنتاج درخت تصمیم نیز برای تحلیل و پیش‌بینی رفتارهای انتخاب مکانی استفاده شده است.بر طبق [4] الگوریتم شبکه عصبی مصنوعی[4] نیز برای انواع مختلفی از مسائل تحلیل مکانی استفاده شده است. استفاده از حسگرهای راه دور یکی از زمینه های اصلیِ نیازمند روش های طبقه بندی است که پیکسل‌های عکس را به دسته‌های برچسب‌دار تقسیم می‌کند.

۲-۱-۲- کاوش قوانین انجمنی مکانی و هم محل ها

بر اساس [5,6,7] قوانین انجمنی مکانی هم مانند کاوش قوانین انجمنی در پایگاه‌داده‌های رابطه‌ای یا تراکنشی، می‌توانند با در نظر گرفتن ویژگی‌های مکانی و پیش‌بینی‌ها، از پایگاه داده‌های مکانی سیئ قوانین بی‌اهمیت و تمرکز روی قوانین جدید و جالب به دانش دامنه نیاز است. کاوش الگوهای هم‌محل مکانی نیز از نظر کلیت مشابه کاوش قوانین انجمنی است منتها از نظر تکنیکی بسیار متفاوت است.

با در اختیار داشتن یک مجموعه داده و ویژگی‌های مکانی و موقعیت های مکانی آن‌ها، یک الگوی هم‌محل نشان‌دهنده زیرمجموعه‌هایی از ویژگی‌هایی است که بسیار نزدیک به هم قرار دارند، بعنوان مثال گونه‌های خاصی از پرندگان که روی نوع خاصی از درختان زندگی می کنند. با این حال یک موقعیت مکانی یک تراکنش نیست و دو ویژگی به ندرت در یک نقطه مکانی قرار دارند. در نتیجه یک محل توصیف‌شده توسط کاربر مورد نیاز است تا بعنوان یک مخزن عمل کرده و اینکه کدام دو ویژگی در یک محله هم محل هستند را بررسی کند. معیارها و الگوریتم‌هایی نیز برای کاوش الگوهای هم محل پیشنهاد شده اند[8,9,10].

۲-۱-۳- خوشه‌بندی مکانی و تحلیل‌الگو

برای درنظرگرفتن اطلاعات مکانی در خوشه بندی ، سه نوع تحلیل خوشه بندی صورت گرفته است:

خوشه‌بندی مکانی : شباهت بین نقاط داده یا خوشه‌ها توسط ویژگی‌های مکانی(مانند موقعیت های مکانی و فاصله) مشخص شده‌اند. خوشه‌بندی مکانی می‌تواند پارتیشنی، سلسله‌مراتبی، مبتنی بر تراکم یا مبتنی بر شبکه باشند.

منطقه‌بندی : منطقه‌بندی یک فرم خاص از خوشه‌بندی است که به منظور بهینه‌سازی یک تابع هدف، اشیاء را در خوشه‌های همجوار مکانی قرار می دهد. بسیاری از کاربردهای جغرافیایی مانند منطقه‌بندی آب‌وهوا، تحلیل چشم‌انداز ، تقسیم‌بندی تصویری بوسیله حسگر از راه دور، معمولا نیازمند این هستند که خوشه‌ها به صورت جغرافیایی همجوار باشند.

تجزیه و تحلیل الگوی نقطه : تحلیل الگوی نقطه‌ای که به تحلیل hot spot نیز معروف است، روی تشخیص جمع‌شدگی غیرمعمول وقایع در فضا متمرکز است(مانند خوشه‌های جغرافیایی بیماری، جرم و جنایت و تصادفات رانندگی) . مساله اصلی تحقیقات، تعیین این است که آیا تعداد نقاط رویداد مشاهده شده بیش از حد معمول برای یک منطقه است یا خیر(به عنوان مثال در یک فاصله معین از یک موقعیت مکانی).

Archive of SID
www.SID.ir

۲-۱-٤- کشف روند

روند اشیاء هنگامیکه ویژگی آن ها با توجه به همسایگی برخی اشیاء مکانی تغییر می‌کند، شناسایی می‌شود. روند مکانی تحت عنوان تغییرِ یک یا چند ویژگی غیرمکانی زمانی که از شیء اولیه فاصله می‌گیرد، تعریف می‌شود. الگوریتم‌های مورد استفاده در این روش کشف روند سراسری[5] و کشف روند محلی[6] هستند.

۳- کاربردهای تحلیل و کاوش داده های مکانی

در این بخش به مرور جدیدترین کارهای انجام شده با تمرکز بر کاربردهای کاوش داده‌های مکانی براساس طبقه بندی ارائه شده در بخش قبل پرداخته شده است.

۳-۱- روش های مبتنی بر طبقه بندی

a. روشی ترکیبی برای کاوش داده‌های مکانی مبتنی بر روابط توپولوژیک فازی و الگوریتم تکاملی MOSES

تصمیم‌گیری با کیفیت بالا در برنامه‌ریزی مکانی استراتژیک، به شدت وابسته به استخراج دانش از حجم وسیعی از داده‌ها است. اگرچه بسیاری از مسائل تصمیم‌گیری مانند توسعه مناطق شهری نیاز به چنین ادراک و استدلالی دارند، روش‌های موجود در این زمینه معمولا از دانشِ عمیق استخراج‌شده از پایگاه داده‌های جغرافیایی غافل می‌شوند و بر اساس روش‌های آماری خالص هستند. با توجه به حجم زیاد داده‌های جمع‌آوری‌شده در پایگاه‌داده‌های مکانی، و عدم قطعیت اشیاء مکانی، کاوش قواعد انجمنی برای نمایش سطحِ بالای دانش، کاری چالش برانگیز است.

تعداد کمی از الگوریتم‌ها، اطلاعات جغرافیایی و غیرجغرافیایی را با استفاده از روابط توپولوژیکی مدیریت می کنند.در [11] یک رویکرد جدید برای کاوش داده‌های مکانی براساس چارچوب تکاملی MOSES ارائه شده است که روش کلاسیکِ برنامه‌نویسی ژنتیک را بهبود می‌بخشد. یک معماری ترکیبی به نام GGeo برای اعمال قوانین کاوش MOSES با در نظر گرفتن روابط توپولوژی فازی از داده‌های مکانی پیشنهاد شده است. به جنبه‌های عدم قطعیت و فازیِ یک مدل غنی شده‌ی روابط توپولوژیکی، با استفاده از محاسبات اتصال ناحیه فازی توجه می‌شود. همچنین، برای غلبه بر مشکل محاسبات بسیار زمانبرِ روابط توپولوژیک فازی، روش جدیدی از پیش پردازش داده‌ها ارائه شده است. GGeo اطلاعات جغرافیایی و غیر جغرافیایی را تحلیل کرده، یاد گرفته، از پارامترهای توپولوژیکی و فاصله استفاده کرده و مجموعه ای از فرمول های محاسباتی–مکانی را به صورت قوانین طبقه‌بندی باز می‌گرداند. رویکرد پیشنهادی نسبت به داده‌های خطادار مقاوم است و تمام مراحل آن به صورت موازی اجرا می شود. این روش می‌تواند در مسائل مختلف مربوط به طبقه‌بندی داده‌های مکانی مورد استفاده قرار بگیرد.

۳-۲- روش‌های مبتنی بر قوانین انجمنی

a. همجوشی داده‌های مکانی در زیرساخت های مکانی مبتنی بر داده‌های پیوندی

با رشد استفاده از ابزارهای همراه، حجم زیادی از داده‌های مکانی تولید شده است که البته به صورت جداگانه جمع‌آوری می‌شود. استفاده از روش‌های همجوشی داده‌ای می تواند منجر به یکپارچگی این منابع داده شده و سرویس‌های بسیار ارزشمندی ایجاد نماید. بدین منظور در [12] از استانداردهای OGC و همچنین استانداردهای وب معنایی استفاده شده است. در مرحله بهبود داده‌ها به دنبال تعیین فاکتورهای کیفیت‌داده می باشد. به عبارتی به دنبال اعمال قوانین اعتبار داده‌ها، پردازش رشته‌ها از نظر صحتِ داده‌های متنی جمع‌آوری شده و ... می باشد. در مرحله هماهنگی داده‌ها به دنبال ایجاد حداکثر همخوانی از نظر ساختار و معنا بین داده‌ها می‌باشد. گام های بعدی نیز براساس این پیش‌پردازش‌های انجام شده به دنبال یافتن ارتباطات بین منابع داده مختلف می باشد. به عبارتی به دنبال ایجاد ارتباط بین مفاهیم مختلفی است که در
منابع داده مختلف تعریف شده اند.

از دیدگاه دیگر، همجوشی داده‌های مکانی نقش مهمی در بازیابی داده‌های مکانی از منابع داده‌ای غیرمتصل (به خصوص در وب) ایفا می‌کند; در نتیجه یک پیش‌شرط برای تصمیم‌گیری سازگار و جامع است. بعنوان مثال، در [13] الگوهای پردازش جغرافیایی برای توصیف زیررال‌های متداول فرایند همجوشی استفاده می‌شوند; در نتیجه ترکیب جریان کاری را پشتیبانی می‌کنند. الگوهای پردازش جغرافیایی، زنجیره‌های فرایندی هستند که معمولا در همجوشی داده‌ها استفاده می‌شوند. پیاده‌سازی بر اساس استانداردهای بازِ ایجاد شده و شامل یک مرورگر ، چندین سرویس پردازش مکانی[7] و یک موتور همجوشی برای پشتیبانی از تلفیق مبتنی بر وب و اجرای جریان کارهای همجوشی داده‌های مکانی به صورت ad hoc است.

b. کاوش کارای الگوهای هم‌محلی مبتنی بر رعایت میزان کارایی ویژگی‌ها در داده‌های مکانی

کاوش سنتی الگوهای هم‌محل تلاش می کند زیرمجموعه‌هایی از ویژگی های مکانی را بیابد که نمونه‌هایی از آن ها در برخی مناطق با هم و به طور متراکم در یک مکان قرار گرفته باشند. در اغلب مطالعات گذشته هم‌محل بودن بعنوان یک معیار جالب در نظر گرفته می شود. مقیاس سنتی ممکن است هم محل‌های جالب را نیابد زیرا کارایی ویژگی‌ها و نمونه‌ها نادیده گرفته شده است. با این حال، در برخی موارد ارزش سودمندی[8] هر نمونه برای کاوش الگوهای مکانی هم‌محل دارای اهمیت بیشتری خواهد بود.

در [14] یک معیار جالب بودنِ جدید برای کاوش الگوریتم‌های هم‌محلِ کارا، از مجموعه داده‌های مکانی با کارایی مختص به آن نمونه ارائه شده است. در معیار جدید جالب بودن، برای اینکه بتوان تاثیر سراسری ویژگی‌های مکان های مشترک را بدست آورد، (intra-utility and inter-utility) در نظر گرفته می شود. یک الگوریتم پایه برای کاوش مکان‌های مشترک ابزارهای کارا ارائه شده است که در آن به منظور کاهش هزینه‌های محاسباتی، استراتژی‌های هرس برای افزایش بهره‌وری معرفی شده اند.

Archive of SID
www.SID.ir

آزمایشات انجام شده روی داده‌های دنیای واقعی و داده‌های مصنوعی نشان می‌دهند که روش پیشنهادی موثر بوده و استراتژی‌های هرس کارا هستند.

c. کاوش افزایشیِ هم‌محلی داده‌های مکانی مبتنی بر کارایی ویژگی‌ها

بر طبق پژوهش [15] در کاوش مکانی هم‌محل‌های بسیار سودمند ، باید سودمندی را بعنوان یک معیار جالب‌بودن در نظر گرفت. این کار باید با درنظر گرفتن مقادیر مختلف یک نمونه خاص که متعلق به ویژگی‌های مختلف است صورت گیرد. هدف الگوریتم UUOC[9]، کاوش افزایشیِ هم‌محل‌های بسیار سودمند روی منابع داده‌های مکانی است که داده‌های این منابع دائما در حال اضافه و ناپدید شدن هستند. زمانی که منبع داده مکانی تغییر می‌کند، کاوش افزایشی هم‌محل‌های بسیار سودمند کار دشواری است، زیرا داده‌های اضافه و یا ناپدیدشده ارتباطات مکانی جدیدی را تولید کرده و ارتباطات قبلی را هم از بین می برند. ارتباطات تغییریافته می‌توانند روی نتایج کاوش هم‌محل‌های بسیار سودمند تاثیر بگذارند. در نتیجه به روزرسانی موثر هم‌محل های‌بسیار سودمند چالش بزرگی است.

الگوریتم UUOC برای به‌روزرسانی موثرِ مقادیر سودمندیِ الگوهای هم‌محل‌های تغییر یافته، پیشنهاد شده است که به‌روزرسانی را با استفاده از نمونه‌های هم‌محل‌های تغییر یافته که فقط با الگوهای تغییر یافته سروکار دارند، انجام می دهد. در این الگوریتم هم محل‌های تغییر یافته به ۴ قسمت تقسیم شده، و یک مفهوم نرخ سودمندی برای قضاوت الگوهای تغییرنیافته معرفی می شود. الگوریتم UUOC کارآمد است و می تواند در دستیابی به نتایج مورد انتظار کمک کند.

d. یافتن بهترین مکانِ قرارگیری برای یک کسب‌وکار با استفاده از کاوش هم‌محلی در داده‌های مکانی-زمانی

برای مشخص‌کردن موقعیت مکانیِ بهینه برای یک کسب‌وکار داده‌شده، باید در نظر داشت که تا زمانی که کسب و کار، مبتنی بر موقعیت مکانیِ حقیقی و مشتریان واقعی باشد، موقعیت مکانی یک فاکتور بسیار مهم برای موفقیت خواهد بود.طبق [16] OptiLocator روشی است که مبتنی بر کاوش هم محل‌های مکانی است و ابعاد زمانی-مکانی سودمند، داده‌های اجتماعیِ مرتبط با فضای شهریِ مورد بررسی را اندازه‌گیری می‌کند. OptiLocator موقعیت مکانی بهینه برای نوع کسب و کار داده شده محاسبه می کند.

OptiLocator شامل مراحل زیر است:

۱. آماده‌سازی و تنظیم داده‌های مکانی-زمانی
۲. خوشه‌بندی داده‌های نشان‌دهنده کسب‌وکارهای موجود با استفاده از DBSCAN
۳. محاسبه شاخص‌های نشان‌دهنده منطقه
۴. محاسبه قوانین هم‌محل‌های مکانی (استفاده از الگوریتم FARICS)
۵. تعیین محل بهینه

در این روش جنبه‌های مالی در نظر گرفته نشده‌اند. و مشکل اصلی فقدان داده‌ی قابل اطمینانی است که بتواند به طور خودکار بررسی‌شود.

۳-۳- روش‌های مبتنی بر خوشه‌بندی

a. خوشه بندی پارتیشنی

روش CLARANS

CLARANS یک روش خوشه‌بندی پارتیشنیِ مبتنی بر الگوریتم k-medoid و جستجوی تصادفی است که در [17] ارائه شده و هدف آن شناسایی ساختارهای مکانیِ ممکن بین داده‌هاست.

نتایج آزمایشات تجربی انجام شده نشان می‌دهد که روش CLARANS نسبت به روش های خوشه‌بندی موجود بسیار کارآمد و موثر است. CLARANS داده‌های اشیاء چند ضلعی را نیز علاوه بر اشیاء نقطه‌ای، به طور کارآمدی خوشه‌بندی می کند.روش CLARANS در مقایسه با PAM، برای مجموعه داده‌های کوچک سریع‌تر است اما هرچه مجموعه داده بزرگتر باشد اختلاف کارایی بیشتر می شود. و در نهایت با استفاده از CLARANS، دو الگوریتم کاوش داده‌های مکانی برای کشف ارتباطات بین ویژگی‌های مکانی و غیرمکانی ارائه شده است. هر دو الگوریتم توانایی کشف دانشی که یافتن آن برای الگوریتم‌های کاوش داده‌های مکانی دیگر مشکل است را دارند.

الگوریتم خوشه بندی توزیع شده D2CA

تکنیک‌های داده‌کاوی توزیع‌شده و به طور کلی خوشه‌بندی توزیع‌شده به طور گسترده استفاده می‌شوند، زیرا با حجم داده‌های بسیار زیاد و ناهمگن که نمی‌توانند به طور مرکزی تجمیع شوند، سروکار دارند.در [18] روش D2CA ارائه شده است که یک روش خوشه‌بندی جدید برای مجموعه داده‌های مکانی بسیار بزرگ ناهمگن و توزیع شده است . این روش مبتنی بر الگوریتم پارتیشنی k-means است اما تعداد خوشه‌های سراسری را به صورت پویا تولید می کند. در واقع این روش از یک فاز تجمیعِ توسعه‌یافته استفاده می کند. فاز تجمیع طوری طراحی شده است که روند کلی در تخصیص زمان و حافظه کارآمد است. نتایج اولیه آزمایشات نشان می دهد که این رویکرد، نتایج با کیفیت بالا تولید می کند و به خوبی قابل توسعه است. الگوریتم D2CA در مقایسه با دو الگوریتم خوشه‌بندی سلسله‌مراتبی BIRCH و CURE بسیار کارآمد است.

b. خوشه بندی سلسله مراتبی

استفاده از خوشه بندی سلسله‌مراتبی موازی برای حل چالش های داده‌های کلان مکانی

خوشه‌بندی می تواند با دسته‌بندی کردن اشیاء مشابه، به مدیریت داده‌های کلان کمک کند. با این حال، روش‌های خوشه‌بندی قادر به خوشه‌بندی مجموعه داده‌های بزرگ نیستند(بعنوان مثال بیشتر از ده میلیون شئ). برای حل این مشکل در [19] روشی ارائه شده است. K-Tree یک ساختار داده و الگوریتم خوشه‌بندی است که می تواند برای مجموعه داده‌های جریانی بزرگ مورد استفاده قرار گیرد. K-Tree روی داده‌های مکانی (تصاویر ماهواره) اعمال شده است و از یک محیط تک‌هسته ای به یک محیط چندهسته ای

Archive of SID
www.SID.ir

گسترش داده شده است. K-Tree توانسته است مجموعه داده‌های بزرگتر را نسبت به رویکردهای پایه، موثرتر خوشه بندی کند.

c. خوشه‌بندی مبتنی بر چگالی

یک چارچوب خوشه‌بندی و تجزیه و تحلیل مبتنی بر چندضلعی برای کاوش مجموعه داده‌های مکانی

چندضلعی‌ها نمایش طبیعی را برای بسیاری از انواع اشیاء جغرافیایی فراهم می‌کنند؛ در نتیجه روش‌های کاوش داده مبتنی بر چندضلعی برای کاوش مجموعه داده‌های جغرافیایی بسیار مفید هستند. در [20] روشی برای کشف چندین مجموعه داده جغرافیایی ارائه شده است که روابط پنهانِ ذاتی دارند. چندضلعی ها ابتدا از چندین مجموعه‌داده نقاط جغرافیایی تولید شده‌اند که این کار با استفاده از یک الگوریتم محدودکننده مبتنی بر چگالی به نام DCONTOUR انجام می شود. سپس از الگوریتم خوشه‌بندی مبتنی بر چگالی به نام Poly-SNN با توابع جدید توازن برای تقسیم چندضلعی‌ها به زیر خوشه‌های چندضلعی استفاده می شود. در نهایت، تکنیک‌های تحلیل پس‌پردازش پیشنهاد شده‌اند تا الگوهای جالب و دانش خلاصه‌ی هدایت‌شده توسط کاربر را از فراخوشه‌ها استخراج کنند. این تکنیک‌ها از توابع plug-in reward استفاده می کنند.

d. خوشه بندی مبتنی بر شبکه

یک رویکرد اکتشافی برای کشفِ آنلاین خوشه‌های مکانی ناشناخته از الگوریتم‌های جریان مبتنی بر شبکه

بر طبق [21]خوشه‌بندی مکانیِ آنلاین جریان‌هایی با موقعیت وسیع، مناسبِ کاربردهای زیادی مانند نظارت بر ترافیک شهری و حوادث عظیم است. برای تشخیص بلادرنگِ این خوشه‌های مکانی می توان از الگوریتم‌های مبتنی بر شبکه استفاده کرد که دامنه مکانی را به سلول‌های گسسته تقسیم می‌کند. مزیت اصلی این رویکرد، این است که به جای مقایسه پرهزینه فاصله در الگوریتم‌های مبتنی بر چگالی، تعداد اشیاء در حال حرکت مربوط به هر سلول را می‌شمارد. اما ممکن است در حین این فرایند الگوریتم نتواند خوشه‌های اشیاء متحرکی که از نظر مکانی و زمانی به سلول‌های مجاور منتقل شده اند را تشخیص دهد. برای حل این مسئله، یک رویکرد اکتشافی مبتنی بر چگالی ارائه شده است که نسبت به اشیاء متحرک مربوط به سلول‌های مجاور حساس است. این رویکرد سلول‌ها را به شکاف‌های داخلی تقسیم می کند. سپس چگالی هر سلول با جمع کردن تعداد اشیاء آن سلول با اشیاء شکاف های داخلی سلول‌های مجاور آن، توسط یک تابع وزنی محاسبه می شود. برای جلوگیری از اثرات جانبی تشخیص خوشه‌های نادرست، روش اکتشافی فقط روی سلول‌های گذرا اعمال می‌شود. سلول‌های گذرا به سلول‌هایی گفته می‌شود که چگالی آن ها کمتر اما نزدیک به مقدار آستانه چگالی باشد.

برای بررسی رابطه بین تعداد خوشه‌های تشخیص داده‌نشده و خوشه‌های اشتباه تشخیص داده شده، رویکرد اکتشافی توسط چندین آستانه گذرا و با استفاده از جریان داده‌های دنیای واقعی و الگوریتم خوشه‌بندی مبتنی بر شبکه DG2CEP در [22] ارزیابی شده است. براساس نتایج آزمایشات مشخص شده است که برای وزن‌های خطی، یک آستانه بین ۵۰ تا ۶۰ درصد از آستانه اصلی، بهترین نتایج را تولید کرده است. مخصوصا برای آستانه گذرا ۶۰ درصد ، تعداد خوشه‌های شناسایی‌نشده از ۸۰ درصد به ۲۳.۵۷ درصد کاهش یافته است. برای وزن‌های نمایی، با کاهش w.r.t تعداد خوشه‌های شناخته نشده و اشتباه شناسایی شده هم کاهش می یابند.

۳-٤- روش های مبتنی بر کشف روند

a. شناسایی کارای روند مسیر حرکت در شبکه های حسگر بی سیم

در [23] برای حل مساله تشخیص توزیع‌شده‌ی کارای پیش‌فرض‌های ذخیره‌کننده روندهای مسیر حرکت اشیاء متحرک ، که با در نظرگرفتن یک محدوده چندضلعی ارزیابی شده، راهکاری ارائه شده است. در این مساله داده‌ها(موقعیت مکانی، زمان) بوسیله شبکه‌های حسگر بی‌سیم[10] ردیابی شده‌اند. به طور خاص، الگوریتم‌های توزیع‌شده درون شبکه ای برای تشخیص دو پیش‌فرض روند حرکت بررسی شده اند؛ حرکت پیوسته به سمت[11] و حرکت مداوم به سمت[12]. ابتدا برای یک شئ واحد و سپس برای انواع متناظر چندین شیء. در واقع تمرکز بر روی تشخیصِ کارای پیش‌فرض‌هایی است که می‌گویند آیا یک شیء داده‌شده به سمت یک منطقه داده‌شده با محدوده ی چندضلعی حرکت می کند یا خیر. ارزیابی الگوریتم‌های CMT و PMT نشان می‌دهد که این روش‌ها صرفه‌جویی قابل ملاحظه‌ای در کاهش تعداد پیام‌های تبادل شده در شبکه نسبت به رویکرد ساده که هر تمام موقعیت مکانی های تشخیص داده‌شده را (همراه برچسب زمانی آن)به sink انتقال می دهد، دارد.

٤- ارزیابی مقایسه‌ای

به منظور ارزیابی مقایسه ای کارهای انجام شده در یک حوزه، ابتدا باید معیارهای مناسبی برای مقایسه تعریف شود. در این مقاله، با مطالعه و بررسی دقیق مقالات در حوزه تحلیل داده های مکانی، ویژگی های مورد تأکید هر یک از مقالات استخراج شده و براساس ویژگی ها، معیارهای مقایسه مطابق با جدول (۱) تهیه شده است. در این جدول، مهمترین ویژگی های کارهای گذشته شامل میزان انطباق با استانداردها، کارایی، میزان جامعیت در برخورد با ویژگی های داده های مکانی و میزان کاربردی بودن روش در کاربردهای عملیاتی را مد نظر قرار گرفته است.

٥- جمع بندی و نتیجه‌گیری

در این مقاله، روش‌های ارائه شده برای تحلیل و کاوش داده های مکانی مورد بررسی قرار گرفته است. با توجه به تفاوت مسائل مورد بررسی در هر یک از مقالات و همچنین استفاده از مجموعه داده‌های متفاوت برای انجام ارزیابی، امکان مقایسه روش‌ها از نظر میزان کارایی با هم وجود نداشت.

مراجع

[1] Ester, M., Kriegel, H. P., & Sander, J. (1997). *Spatial data mining: A database approach. In Advances in spatial databases* (pp. 47-66). Springer Berlin/Heidelberg.

Archive of SID
www.SID.ir

جدول ۱- ارزیابی مقایسه ای روش های موجود

ردیف	نام روش	ویژگی های روش												مرجع
		ارتباط با وب معنایی	بررسی داده‌های ناهمگون	پشتیبانی از کلان داده‌ها	استفاده از متن کاوی برای کاوش داده‌های مکانی	توجه به برچسب های زمانی و سری های زمانی	توجه به جزئیات لازم برای پیاده سازی در مسائل واقعی	استفاده از روش های خوشه بندی و دسته بندی	توجه به ویژگی های هندسی داده‌ها	ارائه مدل محاسباتی یا روش رسمی ارزیابی	استفاده از فرادادهها	قابلیت تنظیم نتایج توسط کاربر	توجه به روابط توپولوژیک داده‌های مکانی	
۱	روشی ترکیبی برای کاوش داده‌های مکانی مبتنی بر روابط توپولوژیک فازی و الگورتیم تکاملی MOSES		✓					✓	✓	✓	✓		✓	[11]
۲	همجوشی داده‌های مکانی در زیرساخت های مکانی مبتنی بر داده‌های پیوندی	✓	✓	✓	✓									[12]
۳	کاوش کارآی الگوهای هم محلی مبتنی بر رعایت میزان کارایی ویژگی ها در داده‌های مکانی							✓		✓		✓		[14]
۴	کاوش افزایشی هم محلی داده‌های مکانی مبتنی بر کارایی ویژگی ها					✓		✓		✓		✓		[15]
۵	یافتن بهترین مکان قرارگیری برای یک کسب و کار با استفاده از کاوش هم محلی در داده‌های مکانی-زمانی					✓	✓		✓					[16]
6	روش CLARANS		✓					✓	✓	✓	✓			[17]
7	الگوریتم خوشه بندی توزیع شده در کاوش داده‌های مکانی		✓	✓			✓	✓		✓				[18]
8	استفاده از خوشه بندی سلسله‌مراتبی موازی برای حل چالش های داده‌های کلان مکانی			✓			✓	✓						[19]
9	یک چارچوب خوشه بندی و تجزیه و تحلیل مبتنی بر چندضلعی برای کاوش مجموعه داده‌های مکانی					✓		✓	✓	✓	✓	✓	✓	[20]
10	یک رویکرد اکتشافی برای کشفِ آنلاین خوشه‌های مکانی ناشناخته از الگوریتم‌های جریان مبتنی بر شبکه			✓		✓		✓						[21]
11	شناسایی کارای روند مسیر حرکت در شبکه های حسگر بی سیم			✓		✓	✓							[23]

Archive of SID

[2] Andrienko, G. L., & Andrienko, N. V. (1999). *Data mining with C4. 5 and interactive cartographic visualization. In User Interfaces to Data Intensive Systems*, 1999. Proceedings (pp. 162-165). IEEE.

[3] Thill, J. C., & Wheeler, A. (2000). *Tree induction of spatial choice behavior. Transportation Research Record*: Journal of the Transportation Research Board, (1719), 250-258.

[4] Cleve, C., Kelly, M., Kearns, F. R., & Moritz, M. (2008). *Classification of the wildland–urban interface: A comparison of pixel-and object-based classifications using high-resolution aerial photography*. Computers, Environment and Urban Systems, 32(4), 317-326.

[5] Appice, A., Ceci, M., Lanza, A., Lisi, F. A., & Malerba, D. (2003). *Discovery of spatial association rules in geo-referenced census data: A relational mining approach*. Intelligent Data Analysis, 7(6), 541-566.

[6] Koperski, K., & Han, J. (1995). *Discovery of spatial association rules in geographic information databases*. In Advances in spatial databases (pp. 47-66). Springer Berlin/Heidelberg.

[7] Mennis, J., & Liu, J. W. (2005). *Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change*. Transactions in GIS, 9(1), 5-17.

[8] Shekhar, S., & Huang, Y. (2001, July). *Discovering spatial co-location patterns: A summary of results*. In Sstd (Vol. 2121, pp. 236-256).

[9] Huang, Y., Pei, J., & Xiong, H. (2006). *Mining co-location patterns with rare events from spatial data sets*. Geoinformatica, 10(3), 239-260.

[10] Lu, Y., & Thill, J. C. (2008). *Cross-scale analysis of cluster correspondence using different operational neighborhoods*. Journal of Geographical Systems, 10(3), 241-261.

[11] Goudarzi, A. H., & Ghadiri, N. (2017). *A hybrid spatial data mining approach based on fuzzy topological relations and MOSES evolutionary algorithm*. arXiv preprint arXiv:1704.06621.

[12] Wiemann, S., & Bernard, L. (2016). *Spatial data fusion in spatial data infrastructures using linked*

www.SID.ir

data. International Journal of Geographical Information Science, 30(4), 613-636.

[13] Wiemann, S. (2017). *Formalization and web-based implementation of spatial data fusion*. Computers & Geosciences, 99, 107-115.

[14] Wang, L., Jiang, W., Chen, H., & Fang, Y. (2017, March). *Efficiently Mining High Utility Co-Location Patterns from Spatial Data Sets with Instance-Specific Utilities*. In International Conference on Database Systems for Advanced Applications (pp. 458-474). Springer, Cham.

[15] Wang, X., & Wang, L. (2017, February). *Incremental mining of high utility co-locations from spatial database*. In Big Data and Smart Computing (BigComp), 2017 IEEE International Conference on (pp. 215-222). IEEE.

[16] Bembenik R., Szwaj J., Protaziuk G. (2017) OptiLocator: *Discovering Optimum Location for a Business Using Spatial Co-location Mining and Spatio-Temporal Data*. In: Kryszkiewicz M., Appice A., Ślęzak D., Rybinski H., Skowron A., Raś Z. (eds) Foundations of Intelligent Systems. ISMIS 2017. Lecture Notes in Computer Science, vol 10352. Springer, Cham

[17] Ng, R. T., & Han, J. (2002). *CLARANS: A method for clustering objects for spatial data mining*. IEEE transactions on knowledge and data engineering, 14(5), 1003-1016.

[18] Bendechache, M., & Kechadi, M. T. (2015, July). *Distributed clustering algorithm for spatial data mining*. In Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2015 2nd IEEE International Conference on (pp. 60-65). IEEE.

[19] Woodley, A., Tang, L. X., Geva, S., Nayak, R., & Chappell, T. (2016, December). *Using parallel hierarchical clustering to address spatial big data challenges*. In Big Data (Big Data), 2016 IEEE International Conference on (pp. 2692-2698). IEEE.

[20] Wang, S., & Eick, C. F. (2014). *A polygon-based clustering and analysis framework for mining spatial datasets*. GeoInformatica, 18(3), 569-594.

[21] Junior, M. R., Endler, M., Casanova, M. A., Lopes, H., & e Silva, F. S. (2016, September). *A Heuristic Approach for On-line Discovery of Unidentified Spatial Clusters from Grid-Based Streaming Algorithms*. In International Conference on Big Data Analytics and Knowledge Discovery (pp. 128-142). Springer International Publishing.

[22] Junior, M. R., Endler, M., & e Silva, F. J. D. S. (2017). *An on-line algorithm for cluster detection of mobile nodes through complex event processing. Information Systems*, 64, 303-320.

[23] Avci, B., Trajcevski, G., Tamassia, R., Scheuermann, P., & Zhou, F. (2017). *Efficient detection of motion-trend predicates in wireless sensor networks*. *Computer Communications*, *101*, 26-43.

[24] Sumathi, N., Geetha, R., & Bama, S. S. (2008). Spatial data mining—techniques trends and its applications. Journal of Computer Applications, 1(4), 28-30.

زیرنویس‌ها

[1] Geographic information system (GIS)
[2] Spatial Data Mining (SDM)
[3] Global Positioning System (GPS)
[4] Artificial Neural Network (ANN)
[5] Global Trend-Detection
[6] Local Trend-Detection
[7] geoprocessing
[8] utility value
[9] Utility Update of Co-locations
[10] wireless sensor network (WSN)
[11] Continuously Moving Towards (CMT)
[12] Persistently Moving Towards (PMT)

Archive of SID
www.SID.ir