

کشف و تحلیل مناطق جذاب با استفاده از تصاویر دارای برچسب مکانی

سعید دهقانی^{۱*}، علی اصغر آل شیخ^۲

^۱ دانشجوی کارشناسی ارشد سیستم‌های اطلاعات مکانی - دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه

نصیرالدین طوسی

saeed_dehghani@email.kntu.ac.ir

^۲ استاد دانشکده‌ی مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه نصیرالدین طوسی

alesheikh@kntu.ac.ir

(تاریخ دریافت دی ۱۳۹۹، تاریخ تصویب اسفند ۱۳۹۹)

چکیده

گسترش فناوری‌های نوین در دستگاه‌های رقمی و ارتباطات باعث محبوبیت و فراگیر شدن شبکه‌های اجتماعی شده است. محتوای دارای برچسب مکانی با افزودن بعد جدیدی به شبکه‌های اجتماعی، امکان ارتباط بین فضای مجازی و واقعی را فراهم کرده‌اند. این تصاویر می‌توانند نشان‌دهنده‌ی تعامل کاربران با محیط باشند؛ به همین دلیل در کاربردهای فراوانی مثل کشف مناطق جذاب، مدیریت و برنامه‌ریزی شهری و توصیه به گردشگران می‌توانند مورد استفاده قرار گیرند. استفاده از تصاویر دارای برچسب مکانی به منظور کشف و تحلیل مناطق جذاب، نسبت به روش‌های سنتی مانند سرشماری و نظرسنجی مزایایی مثل صرفه‌جویی در وقت و هزینه را دارند، از این رو مورد توجه محققان قرار گرفته‌اند. هدف این تحقیق کشف مناطق جذاب با استفاده از داده‌های دارای برچسب مکانی است، به گونه‌ای که از داده‌های توصیفی جهت بهبود خوشه‌های ایجاد شده استفاده گردد و نتایج نسبت به روش‌های رایج پیشین بهبود یابد. همچنین استخراج اطلاعات معنایی مناسب و تحلیل در شرایط مختلف جهت شناخت مناطق جذاب و درک دلیل جذابیت آن‌ها از دیگر اهداف این تحقیق است. در این تحقیق از تصاویر دارای برچسب مکانی فلیکر که از شهر نیویورک، در بازه‌ی زمانی ۲۰۱۵ تا ۲۰۱۸ میلادی اخذ شده بودند، نوفه و افزونگی داده‌ها در مرحله‌ی پیش‌پردازش حذف شد. سپس با استفاده از روش HDBSCAN، داده‌ها خوشه‌بندی شدند و خوشه‌های مجاور که از نظر برچسب‌های متنی مشابه بودند، با یکدیگر ادغام شدند و ۱۰۶ منطقه‌ی جذاب شناسایی شد. سپس سطح مقعری با استفاده از روش α -shape به نقاط برازش شد و اطلاعات معنایی شامل برچسب‌های متمایز، نام و دسته‌بندی برای مناطق جذاب انتخاب شد. آنگاه مناطق جذاب بر اساس نوع کاربران بازدیدکننده، میزان بازدید در شرایط بافت مختلف و احساسات کاربران تحلیل شدند. نتایج حاصل از ارزیابی نشان از کشف مناطق جذاب در اشکال، ابعاد و چگالی‌های مختلف است. این مناطق با ۶۶ درصد از کل جاذبه‌های برتر TripAdvisor مطابقت داشته‌اند، درحالی‌که برای روش DBSCAN ساده ۵۳ درصد مطابقت وجود داشته است. همچنین مناطقی که با جاذبه‌های TripAdvisor هم‌پوشانی داشت‌اند، نام‌گذاری آن‌ها ۷۶ درصد مشابهت داشته است.

واژگان کلیدی: مناطق جذاب، شبکه‌های اجتماعی مکان‌مبنا، تصاویر دارای برچسب مکانی، خوشه‌بندی مکانی، فلیکر

* نویسنده رابط

۱- مقدمه

انقلاب فناوری‌های جدید در حوزه‌ی ارتباطات و دستگاه‌های دیجیتال باعث شده تا حجم عظیمی از داده‌های ناهمگن با سرعت زیادی تولید شوند که به آن‌ها کلان داده^۱ اطلاق می‌گردد. عمده‌ی آن چیزی که امروزه به‌عنوان کلان داده شناخته می‌شود، توسط حسگرهای مختلف تولید می‌گردد. به تمامی منابع تولید کلان داده می‌توان شبکه‌های اجتماعی را هم افزود که در آن‌ها کاربران همانند حسگرهای اجتماعی عمل کرده و بدون این که هدف اصلی آن‌ها باشد، به تولید کلان داده می‌پردازند که این موضوع زمینه‌ساز مطالعه‌ی رفتار کاربران شده است [۱، ۲].

شبکه‌های اجتماعی امروزه به شکل‌های مختلف مانند سرویس‌های اشتراک‌گذاری تصاویر، وب‌نوشت‌ها و سرویس‌های گفتگو بین کاربران به‌شدت فراگیر شده‌اند، به‌گونه‌ای که تا سال ۲۰۱۹ میلادی بیش از ۳/۵ میلیارد کاربر در شبکه‌های اجتماعی فعالیت می‌کنند که برابر با ۴۲ درصد از جمعیت جهان است [۳]. شبکه‌های اجتماعی درواقع ساختاری انتزاعی بین کاربران هستند که از طریق روابطی مثل دوستی یا منافع مشترک به هم متصل هستند. با پیشرفت فناوری‌های تعیین موقعیت و ارتباطات بی‌سیم، بعد مکان به شبکه‌های اجتماعی افزوده شده است که پلی بین جهان فیزیکی و شبکه‌های اجتماعی ایجاد کرده است [۴].

شبکه‌های اجتماعی از دو جهت تحت تأثیر بعد مکان قرار گرفته‌اند؛ گروهی از شبکه‌های اجتماعی، امکان اشتراک‌گذاری محتوای غنی‌شده با برچسب‌های مکانی^۲ را به کاربران ارائه می‌دهند. گروه دیگر نوعی از شبکه‌های اجتماعی هستند که مکان در آن‌ها به‌عنوان یک موجودیت مستقل عمل می‌کند و می‌تواند با دیگر مکان‌ها و کاربران ارتباط داشته باشد که به آن‌ها شبکه‌های اجتماعی مکان‌مبنا^۳ می‌گویند [۴].

شبکه‌های اجتماعی مکان‌مبنا امکانات و کاربردهای مکانی بیشتری را برای کاربران فراهم می‌کنند؛ ولی از طرف دیگر شبکه‌های اجتماعی غنی‌شده با محتوای دارای

برچسب مکانی، تعداد کاربران و محبوبیت بیشتری دارند و اطلاعات مکانی در آن‌ها غنی‌تر هستند. استفاده از داده‌های مکانی خام شبکه‌های اجتماعی غنی‌شده با محتوای دارای برچسب مکانی پیچیدگی بیشتری وجود دارد. استفاده از داده‌های مکانی خام (طول و عرض جغرافیایی) باعث تسهیل تجزیه و تحلیل روندها برای انجام پیش‌بینی می‌شود و امکان شناسایی مناطق جذاب جدید در طول زمان را فراهم می‌نماید [۴، ۵].

از میان انواع محتوای دارای برچسب مکانی که شبکه‌های اجتماعی، امکان اشتراک‌گذاری آن‌ها را برای کاربران فراهم می‌کنند، محتوای تصویری بیشتر نشان‌دهنده‌ی علایق و تعامل کاربران در محیط جغرافیایی است، این در حالی است که سایر محتوای دارای برچسب مکانی کمتر از این قابلیت برخوردار هستند؛ برای مثال ممکن است یک کاربر وب‌نوشتی را با موقعیت فعلی خود ارسال کند که محتوای آن در مورد رویدادهای کشور دیگری است که در چنین حالتی، مکان فعلی کاربر که به وب‌نوشت الصاق می‌شود نشان‌دهنده‌ی علاقه و تعامل کاربر با آنجا نیست [۵، ۶].

تصاویر دارای برچسب مکانی جغرافیا و زمان و محتوای تصویری را به هم متصل می‌کند [۷] و امکان کار با داده‌های مکانی-زمانی با تفکیک‌پذیری بالا و به‌روز را فراهم می‌نماید [۸]. این داده‌ها با وجود این که دارای نوفه^۴ هستند ولی می‌توانند در کاربردهای وسیعی همچون جست‌وجو، تبلیغات، برچسب‌گذاری و سامانه‌های توصیه‌گر استفاده شوند [۹] و یکی از کاربردهای بسیار رایج آن‌ها، کشف و تحلیل مناطق جذاب است؛ زیرا مناطق جذاب موردتوجه کاربران هستند و مکرراً مورد عکس‌برداری قرار می‌گیرند. شناخت مناطق جذاب و نحوه‌ی تعامل مردم با آن‌ها، اطلاعات ارزشمندی جهت برنامه‌ریزی و تحلیل مسائل شهری و گردشگری فراهم می‌نماید، بنابراین تحقیقات مربوط به کشف و توصیه‌ی مناطق جذاب، به‌سرعت تبدیل به یک موضوع موردعلاقه‌ی محققان شده است [۱۰].

کشف و تحلیل مناطق جذاب به‌طور متداول شامل ترکیبی از روش‌های کمی و کیفی است که بر اساس ابزارهایی مانند پرسش‌نامه‌ها، مصاحبه‌ها، نظرسنجی‌ها،

^۱ Big Data^۲ Geotagged^۳ Location Based Social Networks

۴ Noise

کشف و تحلیل مکان‌های جذاب استفاده کرده‌اند [۸، ۱۵، ۱۶].
 عمده‌ی تحقیقات انجام‌شده در این حوزه از داده‌های فلیکر^۳ استفاده کرده‌اند [۲، ۵، ۶، ۱۰، ۱۳، ۱۴، ۱۷].
 سرویس فلیکر به دلیل داشتن API^۴ آزاد، عدم نیاز به اجازه‌ی دسترسی کاربران، نداشتن محدودیت زمانی در اخذ تصاویر، ارائه‌ی سرویس تا زمان حال و سوق دادن کاربران به عکاسی از دیدگاه شخصی، برای انجام تحقیق در این حوزه نسبت به سایر شبکه‌های اجتماعی مشابه، مثل پانورامیو و اینستاگرام برتری دارد.

مکان‌های جذاب کشف‌شده در برخی از تحقیقات به شکل نقاط جذاب ارائه شده‌اند [۲، ۷، ۹، ۱۰، ۱۵، ۱۷] و در برخی دیگر به شکل مناطق جذاب ارائه گشته‌اند [۵، ۶، ۱۴].
 نقاط و مناطق جذاب به مکان‌هایی اشاره دارند که به دلایلی توجه مردم را جلب می‌کنند و بازدیدکننده‌ی زیادی دارند. استفاده از مفهوم منطقه به جای نقطه باعث می‌شود اطلاعات بیشتری در مورد شکل مکان ارائه شود؛ زیرا مناطق جذاب مرز مشخصی را تشکیل می‌دهند و به شکل یک چندضلعی ارائه می‌شوند. همچنین امکان انجام عملکردهای مکانی مثل انجام پرس‌وجو و بازیابی اطلاعات با کارایی بیشتری فراهم می‌شود [۶]. اسپرو و همکاران (۲۰۱۷) جهت ایجاد مناطق جذاب، شبکه‌ی سلولی مربعی شکلی به سطح شهر برازش دادند و سلول‌های مجاور که از نظر برچسب‌های متنی مشابه بودند را ادغام کردند. استفاده از شبکه‌ی سلولی موجب می‌شود شکل مناطق به اجتماع سلول‌ها محدود شوند و اندازه‌ی سلول‌ها تأثیر بسزایی در شکل نهایی مناطق داشته باشند [۵، ۱۴]. در برخی مقالات پس از خوشه‌بندی نقاط، به مجموعه نقاط هر خوشه یک سطح برازش داده‌اند. استفاده از سطح مقعر به جای سطح محدب باعث کاهش نواحی خالی از نقطه می‌شود. از این رو در تحقیق هو و همکاران (۲۰۱۵) از روش‌های chi-shaped برای برازش سطح و ایجاد منطقه‌ی جذاب استفاده شده است [۶]. همچنین در تحقیق کو و همکاران (۲۰۱۸) از روش alpha-shape برای این منظور استفاده شده است. استفاده از این روش باوجود مزایایی که دارد، نیازمند تعیین پارامتر alpha به صورت مناسب است [۵].

دستگاه‌های ثبت موقعیت خودکار، سفرنامه‌های برخط، سالنامه‌های آماری و سایر منابع آماری رسمی است. روش‌های سنتی توانایی محدودی در ارائه‌ی اطلاعات مربوط به مناطق جذاب داشته و در استخراج و به‌روزرسانی آن‌ها، هزینه و زمان زیادی را تحمیل می‌کنند و در بعضی موارد مشکلاتی با حریم شخصی افراد ایجاد می‌نماید [۷، ۱۱].

استفاده از تصاویر دارای برچسب مکانی فرصت بزرگی جهت تکمیل روش‌های سنتی و یا جایگزینی با آن‌ها به‌عنوان رویکردی کارآمد و نوآورانه ایجاد کرده است. این رویکرد با هزینه‌ی کم‌تر، سرعت بیشتر، مقیاس‌پذیری بالاتر و با استفاده از خرد جمعی کاربران، امکان انجام تحلیل‌های مکانی-زمانی پویای شهر را فراهم کرده است. نتایج این تحلیل‌های مکانی برای برنامه‌ریزی، مدیریت و پشتیبانی از تصمیم‌گیری^۱ در حوزه‌های منابع شهری، زیرساخت‌های شهر، حمل‌ونقل و کسب‌وکارها کاربرد دارد که نهایتاً می‌تواند منجر به حرکت به سمت شهر هوشمند و توسعه‌ی پایدار شود. نتایج تحلیل‌های مکانی همچنین می‌تواند در مدیریت و برنامه‌ریزی منابع گردشگری، ارائه‌ی توصیه به گردشگران و کشف قوانین سفر استفاده گردد که منجر به حرکت به‌سوی گردشگری هوشمند می‌شود [۱۰، ۱۲-۱۴].
 به دست آوردن مرز دقیق مناطق جذاب با چالش روبه‌رو است زیرا بسیاری از مناطق جذاب مانند سواحل دریا، دیوار و حد مشخصی ندارند؛ بنابراین با استفاده از روش‌های خوشه‌بندی^۲، مرز طبیعی جاذبه‌های گردشگری می‌تواند استخراج شود که این روش از نظر فعالیت‌های انسانی انتخاب مناسب‌تری است [۱۰]. در ادامه تحقیقات پیشین که در این زمینه انجام شده است، بررسی و تحلیل می‌شوند.

منابع داده‌ی مختلفی برای کشف و توصیه مکان‌های جذاب در تحقیقات استفاده شده است. آلبانا و همکاران (۲۰۱۶) مکان‌های از پیش تعریف‌شده در فوراسکوئر را برای ارائه‌ی توصیه استفاده کرده‌اند؛ سپس آن‌ها مکان‌های تعریف‌شده را با محتوای اینستاگرام غنی کرده‌اند [۴]. پالدینو و همکاران (۲۰۱۵) از تصاویر موجود در sfgeo برای بررسی جاذبه‌ها و جریان‌های گردشگری استفاده کرده‌اند [۱۲]. برخی تحقیقات از تصاویر پانورامیو برای

^۳ Flickr

^۴ Application Programming Interface

^۱ Supported Decision-Making

^۲ Clustering

خوشه‌بندی نقاط را می‌توان مهم‌ترین قسمت کشف نقاط و مناطق جذاب دانست. برای این منظور طیف وسیعی از روش‌ها استفاده شده‌اند. چنگ و همکاران (۲۰۱۱) با روش انتقال میانگین^۱، نقاط جذاب را کشف کردند [۹]. پالدینو و همکاران (۲۰۱۵) و اسپيرو و همکاران (۲۰۱۷) از شبکه سلولی مربعی برای این منظور استفاده کرده‌اند [۱۲، ۱۴]. پنگ و هاوونگ (۲۰۱۷) از روش CFSFDP^۲ که یک روش مبتنی بر کشف قله‌های چگالی است برای کشف نقاط جذاب استفاده کرده‌اند [۱۰]. همچنین کو و همکاران (۲۰۱۸) از روش SO^۳ برای کشف مناطق جذاب استفاده کرده‌اند [۵]. روش DBSCAN^۴ در تحقیقات زیادی استفاده شده است [۲، ۶، ۷، ۱۵]. استفاده از روش‌های مبتنی بر تراکم^۵ مانند DBSCAN به دلیل مزایایی که در این حوزه‌ی تحقیق دارند بسیار رایج هستند. این روش‌ها به‌خوبی شکل‌های نامنظم مناطق جذاب را تشخیص می‌دهند و نیازی به دانش قبلی در مورد تعداد خوشه‌ها ندارند. این در حالی است که استفاده از روش‌هایی مانند انتقال میانگین که مبتنی بر مرکزیت^۶ هستند، به دلیل عدم تشخیص خوشه‌ها با اشکال نامنظم معمولاً کارایی کمتری دارند. با این وجود همچنان به دلیل تغییرات مقیاس و تراکم، روش‌های تراکم‌مبنا با چالش روبه‌رو هستند [۱۳]. در تحقیق عمرانی و آل‌شیخ (۱۳۹۷) از روش OPTICS^۷ استفاده شده که یک روش مبتنی بر تراکم است استفاده شده تا مسئله‌ی تغییرات تراکم تا حد امکان حل شود.

استفاده از مکان‌های جذاب کشف‌شده برای ایجاد سیستم توصیه‌گر، موضوع تحقیقات زیادی بوده است [۴، ۷، ۹، ۱۷]. در این دسته از تحقیقات هدف اصلی بررسی شباهت الگوها و علایق کاربران برای ارائه‌ی توصیه‌ی مکان‌ها بوده است. این تحقیقات به‌طور کلی تمرکز کمتری برای استخراج شکل مناطق، کشف اطلاعات معنایی و تحلیل مکان‌های جذاب داشته‌اند. تعدادی از تحقیقات با تمرکز به روش‌های زمین‌آمار به تحلیل مناطق شهری پرداخته‌اند [۸، ۱۶]. آن‌ها با برازش یک شبکه‌ی

سلولی به سطح شهر با استفاده از تراکم هسته^۸، شاخص موران محلی^۹ و گتیس اورد جی^{۱۰} به تحلیل زمین‌آمار پرداخته‌اند. این تحقیقات به کشف اطلاعات معنایی و شکل مناطق نپرداخته‌اند.

از دیگر حوزه‌هایی موردتوجه محققان، استخراج اطلاعات معنایی^{۱۱} به‌صورت خودکار برای مکان‌ها است. این اطلاعات جهت شناخت مناطق جذاب و درک دلیل جذابیت آن ارائه می‌شود. استخراج نام و برچسب‌های متمایز با استفاده از وزن دهی برچسب‌های متنی به روش TF-IDF^{۱۲} روشی است که در تحقیقات زیادی استفاده شده است [۶، ۱۰، ۱۴]. کو و همکاران (۲۰۱۸) از روش TF-IDF به‌صورت هموارسازی شده برای نام‌گذاری استفاده کرده‌اند [۵]. در برخی تحقیقات در کنار برچسب‌های متنی از یک منبع اطلاعاتی اضافه نیز استفاده کردند تا نام و دسته‌بندی برای مکان‌ها ارائه نمایند [۱۵، ۱۷، ۱۸]. گیگلايو و همکاران (۲۰۱۹) از کلاس‌بندی تصاویر برای ارائه‌ی دسته‌بندی مکان‌ها استفاده کرده‌اند [۲]. همچنین هو و همکاران (۲۰۱۵) در کنار برچسب‌های متمایز به ارائه‌ی تصاویر شاخص از مناطق پرداخته‌اند [۶].

هدف از تحقیق پیش رو، کشف مناطق جذاب با استفاده از داده‌های دارای برچسب مکانی است. در این روش خوشه‌های ایجادشده با استفاده از اطلاعات توصیفی بهبود داده می‌شوند؛ به‌گونه‌ای که کارایی مناسبی نسبت به روش‌های رایج پیشین ارائه نماید و توانایی کشف مناطق با تراکم و ابعاد متفاوت را داشته باشد. برای مناطق جذاب کشف‌شده، با ایجاد تغییرات در روش‌های پیشین، اطلاعات معنایی شامل برچسب‌های متمایز، نام و دسته‌بندی ارائه شوند. سپس با ایجاد بهبود در روش‌های قبلی، کاربران گردشگر از شهروندان تشخیص داده شده و مناطق برحسب نوع بازدیدکنندگان و در شرایط بافتی مختلف تحلیل شوند. در نهایت احساسات کاربران در مورد مناطق جذاب، با استفاده از نظرات آن‌ها و روش‌های تحلیل احساسات^{۱۳} بررسی می‌شوند.

^۸ Kernel Density

^۹ Local Moran's I

^{۱۰} Getis-Ord General

^{۱۱} Semantic

^{۱۲} Term Frequency - Inverse Document Frequency

^{۱۳} Sentiment Analysis

^۱ Mean Shift

^۲ Clustering by Fast Search and Find of Density Peaks

^۳ Spatial Overlap

^۴ Density-Based Spatial Clustering of Applications with Noise

^۵ Density-based

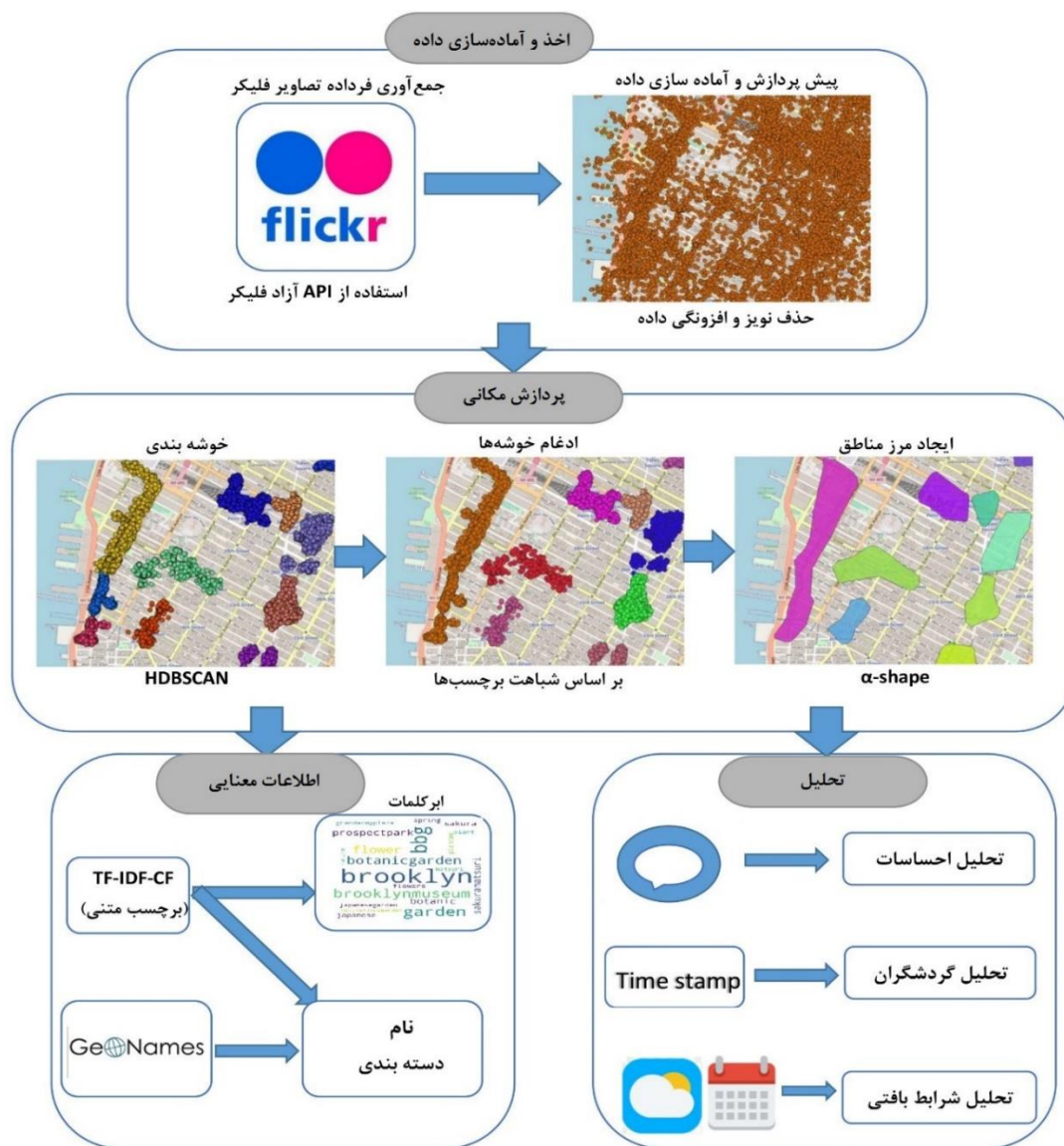
^۶ Centroid-based

^۷ Ordering Points To Identify the Clustering Structure

۲- روش تحقیق

روش ارائه شده در این تحقیق در سه گام اصلی انجام می شود. در گام اول فراداده‌ی تصاویر اخذ می شود و با انجام پیش پردازش، نوفه ها تا حد امکان کاهش داده می شوند. در گام دوم خوشه بندی مکانی از طریق روش HDBSCAN که ترکیبی از روش های سلسله مراتبی و تراکم مبنا است، انجام می گیرد. سپس خوشه های مجاور

که شباهت زیادی دارند از طریق برچسب ها شناسایی و ادغام می شوند. در گام سوم اطلاعات معنایی برای خوشه ها استخراج می شود و به مناطق کشف شده معنی داده می شود. پس از انجام این مراحل، مناطق جذاب کشف شده در شرایط بافت مختلف و از دیدگاه شهروندان و گردشگران، تحلیل می شوند. روند انجام تحقیق در تصویر ۱ آورده شده است.



تصویر ۱- روند نمای انجام تحقیق

۲-۱- جمع آوری اطلاعات

داده های فلیکر با دقت مکانی بالا، بدون نیاز به اجازه ی دسترسی کاربران، به صورت رایگان و بدون محدودیت زمانی تا زمان حال در دسترس هستند. اخذ فراداده ی

تصاویر با انجام پرس و جو^۱ در API فلیکر و با تعیین پارامترهای محدوده ی زمانی و محدوده ی مکانی مورد نظر انجام می شود. فرمت خروجی داده ها json در نظر گرفته

^۱ Query

شده است. همچنین با توجه به محدودیتی که از جانب فلیکر برای درخواست‌ها در نظر گرفته شده است، در هر پرس‌وجو تنها ۴۰۰۰ رکورد داده بازگردانده می‌شود که در صفحات جداگانه قرار می‌گیرند و در هر صفحه نهایتاً ۵۰۰ رکورد نمایش داده می‌شود، بنابراین ارسال درخواست‌ها در بازه‌های زمانی و محدوده‌های جغرافیایی کوچک‌تر انجام می‌شود [۱۹]. به‌منظور دریافت نظرات هر تصویر، با استفاده از API مربوطه و با ارائه‌ی شناسه‌ی یکتای تصویر و فرمت خروجی داده، مجموعه نظرات مربوط به آن تصویر بازگردانده می‌شود. نمونه‌ی فراداده‌ی یک تصویر در جدول ۱ آمده است.

جدول ۱- نمونه‌ای از فراداده‌ی یک تصویر

پارامترها	مقادیر
شناسه‌ی تصویر	۱۶۷۲۲۸۷۴۰۶۶
شناسه‌ی کاربر	N00@۴۹۱۳۵۱۴۴
نام کاربر	andy in nyc
موقعیت اخذ عکس	۴۰/۵۷۹۵۶۶ -۷۴/۰۷۵۷۲۵
زمان اخذ عکس	PM ۵:۰۱:۰۸ ۲۰۱۵/۷/۳
دقت تعیین موقعیت مکانی	Level ۱۶
آدرس فایل تصویری	https://farm9.staticflickr.com/8561/16722874066_72f785118e.jpg
توصیفات	'South Beach, Staten Island NY'
برچسب‌ها	Nyc; newyorkcity; winter; sky; snow; statenisland; southbeach

۲-۲- پیش‌پردازش داده‌ها

داده‌های دارای برچسب مکانی می‌توانند با نوفه همراه باشند. از این‌رو می‌بایست قبل از انجام مراحل بعدی، در یک فرآیند پیش‌پردازش، نوفه‌ها و همچنین افزونگی داده‌ها حذف‌شده یا کاهش یابند. عوامل ایجادکننده‌ی نوفه به همراه روش حذف آن‌ها در زیر آورده شده است [۱۰، ۱۳]:

- در هنگام تقسیم‌بندی مناطق جهت اخذ داده‌ها، ممکن است بعضی از تصاویر بر روی مرز مناطق تقسیم‌شده قرار بگیرند و پس از ادغام داده‌های مناطق، رکوردهای تکراری ایجاد کنند؛ بنابراین بایستی با استفاده از شناسه‌ی یکتای تصاویر، رکوردهای تکراری حذف شوند.
- در هنگام ثبت موقعیت مکانی توسط دوربین یا تلفن همراه، دقت تعیین موقعیت نیز در فراداده‌ی آن‌ها ثبت می‌شود، بنابراین می‌توان تصاویری که موقعیت مکانی ثبت‌شده‌ی آن‌ها دقت کمی دارد، حذف کرد؛

بنابراین تصاویری که پارامتر دقت تعیین موقعیت در آن‌ها از حد آستانه کمتر است، حذف می‌شوند.

- ممکن است یک کاربر تعدادی تصویر از یک مکان در یک محدوده‌ی زمانی مشخص اخذ و بارگذاری کند که باعث ایجاد افزونگی داده و کاهش صحت در مراحل بعدی می‌شود؛ بنابراین تصاویر اخذشده توسط یک کاربر در بازه‌ی زمانی محدود حذف می‌شوند.
- تصاویر مربوط به مجالس شخصی که نشان‌دهنده‌ی تعامل با محیط نیستند و موضوع اصلی‌شان اشخاص و رویداد شخصی است از طریق برچسب‌های متنی با محتوای مربوط به تولد، عروسی و مهمانی شناسایی و حذف می‌شوند.

۲-۳- خوشه‌بندی مکانی

خوشه‌بندی به‌طور کلی شامل قرار دادن اشیاء مشابه در یک گروه است، پس با خوشه‌بندی نقاط اخذ تصاویر می‌توان به کشف مناطق جذاب پرداخت. روش‌های مختلفی برای این منظور وجود دارند، مثل روش‌های مبتنی بر مرکزیت، مبتنی بر شبکه^۱ و سلسله‌مراتبی^۲ که در این بین، روش‌های مبتنی بر تراکم مانند DBSCAN در زمینه‌ی کشف مناطق جذاب بسیار بیشتر موردتوجه بوده و نتایج بهتری ارائه کرده‌اند. این روش‌ها مزایای زیر را نسبت به سایر روش‌ها دارند [۶]:

- توانایی استخراج اشکال نامنظم و طبیعی مناطق جذاب را دارند.
 - نیاز به داشتن دانش قبلی در مورد تعداد خوشه ندارند و همچنین وابستگی به تعداد خوشه‌ها ندارند.
 - در مقابل داده‌های پرت بهتر عمل می‌کنند که به دلیل وجود نوفه زیاد در داده‌های فلیکر، دارای اهمیت است.
 - مبنای ایجاد خوشه، تراکم نقاط است که با جذابیت مناطق در ارتباط است.
- روش DBSCAN با دو پارامتر شعاع همسایگی و حداقل نقاط در هر خوشه، عملیات خوشه‌بندی را انجام می‌دهد که باوجود مزایای ذکرشده، در مقابل تغییرات مقیاس و تراکم، کارایی کمتری دارد. به همین منظور

^۱ Grid-based

^۲ Hierarchy

۴-۲- ادغام خوشه‌ها

با تنظیم پارامتر حداقل نقاط خوشه، می‌توان به خوشه‌های بزرگ‌تر یا کوچک‌تری دست یافت و هرچه مقدار کم‌تری تنظیم شود خوشه‌ها در مقیاس کوچک‌تر کشف می‌شوند. در بعضی موارد به دلایلی مثل محدوده‌ی جغرافیایی بزرگ بعضی از مکان‌ها و یا تراکم کم داده‌های جغرافیایی، باعث می‌شوند یک مکان به چند خوشه تقسیم گردد. برای حل این مسئله ادغام بعضی از خوشه‌ها به‌عنوان یک راه‌حل بیان می‌گردد [۱۰]. برای این منظور شباهت برچسب‌های متنی تصاویر هر خوشه با خوشه‌های مجاور آن بررسی می‌شود و در صورتی که این شباهت از مقداری بیشتر بود این دو خوشه ادغام می‌گردند.

در گام اول برچسب‌های متنی تصاویر هر خوشه با استفاده از روش TF-IDF-CF^۴ بر اساس میزان متمایز بودنشان وزن دهی می‌شوند که در بخش ۲-۶-۱ به‌طور کامل توضیح داده‌شده است. سپس در گام بعد برچسب‌های هر خوشه در یک بردار قرار می‌گیرند و شباهت این دو بردار از طریق تابع شباهت کسینوسی^۵ [۱۰] که مقداری بین ۰ تا ۱ ارائه می‌دهد از طریق رابطه‌ی ۱ محاسبه می‌شود.

$$\text{Sim}(X,Y) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \cdot \|\vec{y}\|} = \frac{\sum_{i=1}^n x_i \times y_i}{\sqrt{\sum_{i=1}^n (x_i)^2} \times \sqrt{\sum_{i=1}^n (y_i)^2}} \quad (1)$$

که در آن X و Y بیانگر دو خوشه‌ای هستند که مقایسه می‌شوند، $\vec{x}$ و $\vec{y}$ بیانگر بردار برچسب‌های دو خوشه‌ی X و Y است و x_i و y_i نشان‌دهنده‌ی وزن برچسب t_i است که در بردارهای $\vec{x}$ و $\vec{y}$ به ترتیب آمده‌اند. با مقایسه‌ی خوشه‌های مجاور در صورتی که شباهت به‌دست‌آمده از مقدار مشخصی بیشتر باشد، آن دو خوشه ادغام می‌شوند.

۴-۵- ایجاد مرز مناطق جذاب

در این مرحله با استفاده از موقعیت تصاویر خوشه‌بندی شده، چندضلعی‌ها ایجاد می‌شوند تا مرز مناطق جذاب را بیان کنند. پوشش محدب^۶ یکی از روش‌های رایج برای این منظور است که حداقل سطح

توسعه‌هایی بر روی روش DBSCAN انجام‌شده است که می‌توان به روش‌های OPTICS و HDBSCAN اشاره کرد که هر دو آن‌ها درواقع ترکیبی از یک روش سلسله مراتبی با یک روش تراکم‌مبنا هستند [۲۰]. با این وجود روش HDBSCAN نسبت به روش OPTICS از نظر مفهوم و ساختار الگوریتمی بهبودیافته است و سرعت بیشتر و عملکرد بهتری دارد. در روش HDBSCAN تبدیل الگوریتم DBSCAN به یک روش خوشه‌بندی سلسله‌مراتبی، نیاز به تعریف حداقل شعاع همسایگی رفع شده و تنها با تعیین پارامتر حداقل نقاط هر خوشه، عملیات خوشه‌بندی قابل انجام است [۲۰]. جزئیات روش HDBSCAN در ادامه آورده شده است.

ورودی: مجموعه داده‌ی نقاط و پارامتر حداقل نقطه‌ی خوشه‌ها

خروجی: خوشه‌بندی HDBSCAN

۱- فاصله اصلی بین تمام اشیا موجود در مجموعه داده محاسبه شود.

۲- درخت پوشای کمینه^۱ برای گراف قابلیت دسترسی متقابل^۲ محاسبه شود و $G_{m_{reach}}$ نامیده شود.

۳- درخت پوشای کمینه گسترش داده و MST_{ext} نامیده شود. سپس با اضافه کردن یک یال حلقه^۳ برای هر رأس، وزنی برابر با فاصله‌ی اصلی آن در نظر گرفته شود.

۴- سلسله‌مراتب HDBSCAN به‌عنوان یک نمودار درختی سلسله مراتبی از MST_{ext} استخراج شود.

a. به تمامی اشیا برچسب مشابهی الصاق می‌شود بنابراین ریشه‌ی درخت تشکیل می‌شود.

b. به ترتیب کاهش وزن، یال‌های تکراری از MST_{ext} حذف شوند.

i. یال‌هایی که وزن یکسان دارند به‌طور هم‌زمان حذف می‌شوند.

ii. پس از حذف یک یال، برچسب‌ها به اجزای متصلی که شامل یک رأس یال حذف‌شده است، الصاق می‌شوند. اگر اشیا حداقل یک یال داشته باشند، به آن‌ها برچسب خوشه‌ی جدید الصاق می‌شود در غیر این صورت به‌عنوان نوفه شناخته می‌شوند.

^۴ Term Frequency - Inverse Document Frequency - Class

Frequency

^۵ Cosine similarity

^۶ Convex Hull

^۱ Minimum Spanning Tree

^۲ Mutual Reachability Graph

^۳ Self loop

محدبی است که شامل مجموعه‌ی نقاط می‌شود. این روش با وجود ارائه‌ی نتایج منحصر به فرد و کارایی بالا از نظر محاسباتی، ممکن است مناطق خالی و بزرگی را ایجاد کند که توسط نقاط اشغال نشده‌اند [۶]. ایجاد یک پوشش مقعر^۱ به عنوان یک راه حل برای کاهش مناطق خالی پیشنهاد شده است [۵، ۶، ۲۱]. الگوریتم α -shape می‌تواند یک سطح را به شکل انعطاف پذیری به مجموعه‌ی نقاط برازش دهد. برای این منظور پارامتر α تعریف شده است که اگر مقدار آن صفر در نظر گرفته شود، خروجی همان پوشش محدب می‌شود ولی در مقادیر بزرگتر، سطح به شکل مقعر نزدیک می‌شود. با این حال اگر مقدار α بزرگ در نظر گرفته شود ممکن است سطح برازش داده شده دندانه دار و پرچین^۲ شود که به همین دلیل باید در انتخاب مقدار این پارامتر دقت شود [۵].

۲-۶- استخراج اطلاعات معنایی

در مراحل قبل چندضلعی هر منطقه‌ی جذاب ساخته می‌شود که محدوده و موقعیت آن را نشان می‌دهد ولی دلیل جذب مردم به این مناطق همچنان نامعلوم است. در این قسمت اطلاعات معنایی با استفاده از برجسب‌های متنی، برجسب‌های زمانی و نظرات کاربران استخراج می‌شوند.

۲-۶-۱- استخراج برجسب‌های متمایز

در هر خوشه برجسب‌های متنی زیادی توسط کاربران برای بیان منطقه‌ی مورد نظر استفاده می‌شوند که اطلاعات ارزشمندی در مورد آن ارائه می‌دهند. این در حالی است که تعداد محدودی از این برجسب‌ها متمایز و خاص آن منطقه است [۶]. برای وزن دهی به میزان متمایز بودن هر برجسب در هر خوشه از روش TF-IDF-CF استفاده شده که نسخه‌ی توسعه یافته روش رایج TF-IDF است که به طور گسترده در تحقیقات انجام شده در این زمینه مورد استفاده قرار گرفته است [۵، ۶، ۱۷]. این روش جهت نشان دادن اهمیت برجسب در خوشه استفاده شده است. برای این منظور در کنار تأثیر تکرار برجسب‌ها در هر خوشه، از تعداد خوشه‌هایی که شامل این برجسب هستند و تعداد کاربرانی که این برجسب را به کار برده‌اند نیز برای

ایجاد تعدیل استفاده می‌شود؛ زیرا برجسبی که در تعداد زیادی از خوشه‌ها به کار برده شود متمایز آن خوشه نیست و همچنین ممکن است یک کاربر خاص از یک برجسب به تعداد زیاد استفاده کند که در این صورت برجسب اهمیت کمتری دارد [۶]. وزن متمایز بودن هر برجسب در هر خوشه، از طریق رابطه‌ی ۲ تعیین می‌شود.

$$W_{ij} = tf_{ij} \times \log \frac{N}{df_j} \times \frac{n_{cij}}{N_{ci}} \quad (2)$$

که در آن W_{ij} وزن برجسب t_i در خوشه‌ی j ، N تعداد کل خوشه‌ها، tf_{ij} تعداد تکرار برجسب t_i در خوشه‌ی j ، df_j تعداد خوشه‌هایی است که برجسب t_i در آن وجود دارد. همچنین n_{cij} تعداد کاربرانی است که برجسب t_i در خوشه‌ی j استفاده کرده‌اند و N_{ci} تعداد کل کاربران در خوشه‌ی j را بیان می‌کند [۲۲]. با استفاده از این روش ده برجسب متمایز با بیشترین وزن برای هر خوشه انتخاب شده است که نشان دهنده‌ی جذابیت‌ها و فعالیت‌های مورد توجه در آن منطقه است.

۲-۶-۲- استخراج نام و دسته‌بندی

به منظور توصیف مناطق جذاب، نیاز به پیدا کردن نام و نوع دسته‌بندی آن‌ها است که برای این منظور در کنار استفاده از برجسب‌های متنی، از نقاط جذاب GeoName^۳ به صورت ترکیبی استفاده شده است. GeoName پایگاه داده‌ای غنی شامل ۲۵ میلیون نقاط جذاب به همراه نام، دسته‌بندی و سایر اطلاعات توصیفی از سراسر جهان و با دسترسی رایگان است.

برای انتخاب نام، از روشی که عبدالمجید و همکاران [۱۸] معرفی کردند با ایجاد تغییراتی استفاده شده است. این تغییرات بدین شکل است که به جای ایجاد بافر حول مرکز هر خوشه، از چندضلعی ایجاد شده در مراحل قبل، جهت جداسازی نقاط GeoName برای هر خوشه استفاده شده است. همچنین از دسته‌بندی‌های معرفی شده در نقاط GeoName و تقسیم آن‌ها به دو دسته‌ی مرتبط و غیر مرتبط استفاده شده است. این روش به این شکل عمل می‌کند که ابتدا بین چندضلعی ایجاد شده برای هر منطقه‌ی جذاب و نقاط GeoName تقاطع مکانی^۴ انجام

^۳ <http://www.geonames.org>

^۴ Spatial Join

^۱ Concave Hull

^۲ Crispy

۲-۷-۲- تحلیل کاربران گردشگر و محلی

با کشف این که کاربران گردشگر یا شهروند هستند، می توان دریافت جذابیت هر منطقه ی جذاب بیشتر مورد توجه کدام دسته بوده است. مثلاً ممکن است یک مرکز خرید کمتر از یک بنای تاریخی مورد توجه گردشگران باشد. برای این منظور روشی در تحقیقات پالومارس و همکاران [۸]، انسالادا و همکاران [۱۶] و مو و همکاران [۱۳] ارائه شده که در آن اگر کاربر فاصله ی بین زمان اخذ اولین آخرین تصویر کمتر از یک ماه، در مدت زمان مورد مطالعه باشد، کاربر به عنوان گردشگر شناخته می شود در غیر این صورت شهروند تشخیص داده می شود. ولی ممکن است گردشگران به خصوص گردشگران داخلی، چندین سفر در طی چند سال به منطقه انجام دهند که با روش ذکر شده، شهروند شناسایی می شوند. برای این منظور، در این تحقیق تغییراتی در این روش ایجاد شده است.

در این تحقیق فرض شده هر گردشگر می تواند چندین سفر کرده باشد به شرطی که هر سفر کمتر از یک ماه به طول بیانجامد و فاصله ی بین هر دو سفر بیشتر از ۶ ماه باشد. همچنین ممکن است برخی شهروندان که فعالیت کمی دارند، به فواصل زمانی طولانی فعالیت کنند. در این صورت ممکن است به اشتباه گردشگر تشخیص داده شوند، به همین دلیل شرط دیگری نیز فرض شده که هر گردشگر در حین سفر خود حداقل ۵ تصویر از منطقه اخذ کرده باشد.

۲-۷-۳- تحلیل در شرایط بافت مختلف

تأثیر زمان و شرایط بافتی بر تعداد بازدیدکنندگان، امری بدیهی است. با بررسی آن ها می توان به الگوهای بازدید پی برد [۱۰]. در این تحقیق مناطق جذاب در شرایط ساعات روز، فصل های سال، شرایط آب و هوایی و دمای هوا بررسی شده است. در این بررسی، ساعات روز به چهار بازه ی زمانی مساوی تقسیم شده است و شرایط آب و هوایی به سه دسته ی صاف یا ابری، مه آلود و بارش برف یا باران تقسیم شده است. همچنین دمای هوا به پنج بازه تقسیم شده است و بازه های آن به این صورت است: کمتر از منفی ۲ درجه سلسیوس (خیلی سرد)، منفی ۲ تا ۸ درجه (سرد)، ۸ تا ۱۸ درجه (معتدل)، ۱۸ تا ۲۸ درجه (نسبتاً گرم) و بیشتر از ۲۸ درجه (خیلی گرم).

می گیرد تا مشخص شود در هر منطقه ی جذاب کدام نقاط GeoName قرار می گیرند. سپس نقاطی که در دسته بندی مرتبط قرار دارند و نام آن ها با برچسب های متنی شباهت دارند در مجموعه ی P قرار می گیرند سپس نام مناطق جذاب به شکل زیر انتخاب می شود:

- اگر هیچ نقطه ای در مجموعه ی P قرار نگیرد، برچسب متنی که بالاترین میزان وزن را در روش TF-IDF-CF بگیرد به عنوان نام انتخاب می شود.
- اگر تنها یک نقطه در مجموعه ی P قرار گرفت، نام و دسته بندی آن نقطه انتخاب می شود.
- اگر چندین نقطه در مجموعه ی P قرار گرفتند، نام و دسته بندی نقطه ای که به مرکز ثقل خوشه از لحاظ فاصله ی اقلیدسی نزدیک تر است انتخاب می شود.

۲-۷-۴- تحلیل مناطق جذاب

۲-۷-۱- تحلیل احساسات کاربران

نظرات در فلیکر عبارات متنی هستند که توسط کاربران برای بیان عقیده و احساس خود راجع به تصاویر، استفاده می شوند و می توانند مفهوم منفی یا مثبت در بر داشته باشند. این منبع غنی اطلاعات در تحلیل مناطق جذاب کمتر مورد توجه قرار گرفته اند.

تحلیل احساسات یا عقیده کاوی، به کشف جهت گیری و شدت احساسات در متون می پردازد و به طور گسترده در بررسی نظرات مشتریان در مورد محصولات استفاده می شود [۲۳، ۲۴]. معمولاً روش های تحلیل عقاید به دو دسته ی مبتنی بر قواعد^۱ و مبتنی بر موارد^۲ تقسیم می شوند. در روش مبتنی بر قواعد، با تعریف قواعد ساده ای توسط روش هایی مثل برچسب گذاری اجزای جمله، به تحلیل عقاید پرداخته می شود. در حالی که در روش مبتنی بر موارد، با استفاده از یادگیری ماشین و آموزش مدل با نمونه های آموزشی تحلیل عقاید انجام می شود [۲۳، ۲۴]. در این تحقیق با استفاده از بسته ی نرم افزاری TextBlob از طریق برچسب گذاری اجزای جملات و استفاده از مجموعه داده ای از معانی کلمات که روشی مبتنی بر قواعد محسوب می شود، تحلیل عقاید انجام گرفته است.

^۱ Rule-based

^۲ Case-based

۲-۸- روش ارزیابی

با مقایسه‌ی موقعیت جاذبه‌های معرفی‌شده در TripAdvisor با موقعیت مناطق جذاب کشف‌شده در این تحقیق، به بررسی کیفیت کشف مناطق جذاب، با استفاده از شاخص‌های Precision، Recall و F1 پرداخته شده است. شاخص Precision درواقع نسبت مناطق جذاب بازیابی شده و مرتبط در تحقیق، به کل مناطق جذاب بازیابی شده در تحقیق است و Recall نسبت مناطق جذاب بازیابی شده و مرتبط در تحقیق، به کل جاذبه‌های اخذ شده در TripAdvisor است و F1 درواقع میانگینی از دو شاخص یاد شده است که از روابط ۴ تا ۶ به دست می‌آیند.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (۴)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (۵)$$

$$F1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \quad (۶)$$

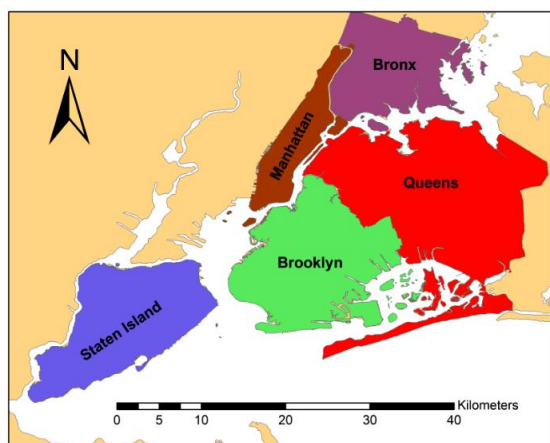
در این روابط، TP تعداد مناطق جذاب کشف‌شده و مرتبط و FP تعداد مناطق جذاب کشف‌شده و غیر مرتبط و FN تعداد جاذبه‌های مرتبط کشف نشده است [۲۵].

۳- پیاده‌سازی و ارزیابی

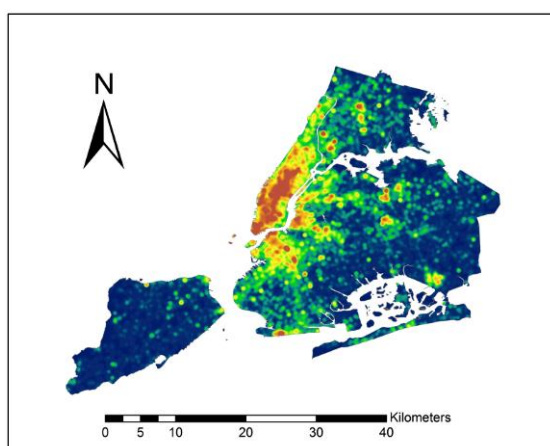
۳-۱- اخذ و آماده‌سازی داده

در این تحقیق شهر نیویورک^۱ در ایالات متحده آمریکا با پنج بخش شهری مورد مطالعه قرار گرفته است و در تصویر ۲ منطقه‌ی مورد مطالعه و بخش‌های مختلف آن نشان داده شده است. این شهر به‌عنوان پرجمعیت‌ترین و متراکم‌ترین شهر ایالات متحده، دارای جاذبه‌های گردشگری شناخته‌شده در سطح جهان و در رتبه‌ی هشتم بیشترین گردشگر در سطح جهان قرار دارد [۲۶]. همچنین در نیویورک حجم قابل‌توجه داده‌های فلیکر وجود دارد و زبان انگلیسی به‌طور گسترده استفاده می‌شود و به همین جهت برای انجام تحقیق مناسب تشخیص داده شده است.

ابتدا فراداده‌ی ۶۲۲۵۱۹ تصویر از سال‌های ۲۰۱۵ تا ۲۰۱۸ میلادی از شهر نیویورک با استفاده از API فلیکر اخذ شدند که با استفاده از کتابخانه‌ی Pandas در پایتون^۲، فایل‌ها از فرمت JSON^۳ به فرمت CSV^۴ تبدیل‌شده و با یکدیگر ادغام شدند تا به یک فایل یکپارچه تبدیل شوند. داده‌های اخذ شده به صورت نقشه گرمایی در تصویر ۳ نشان داده شده است. همچنین داده‌های مربوط به آب‌وهوای شهر نیویورک از طریق وب‌سرویس دانشگاه وایومینگ^۵ و با استفاده از یک خزنده‌ی وب^۶ که با کتابخانه‌ی BeautifulSoup در پایتون ایجاد شده، اخذ شده‌اند که شامل اطلاعات دما، رطوبت، باد و وضعیت آب‌وهوا هستند.



تصویر ۲- بخش‌های پنج‌گانه‌ی شهر نیویورک



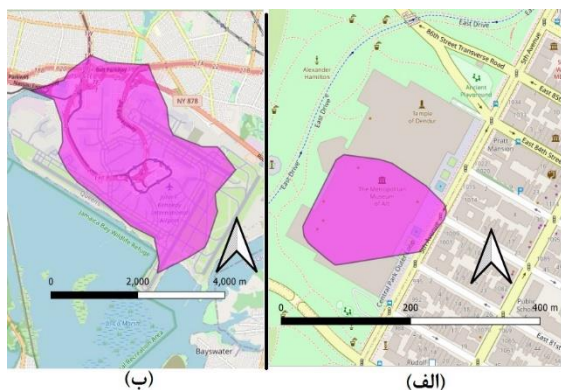
تصویر ۳- نمایش توزیع تصاویر اخذ‌شده در شهر نیویورک با نقشه‌ی گرمایی

^۲ Python^۳ JavaScript Object Notation^۴ Comma-Separated values^۵ University of Wyoming^۶ Web Crawler^۱ New York City

منطقه‌ی جذاب ایجادشده در تصویر ۵ نشان داده شده است. همچنین تفاوت ناحیه خالی ایجادشده در براش سطح مقعر و سطح محدب در تصویر ۶ مقایسه شده است.



تصویر ۴- (الف) خوشه‌ی ایجادشده پس از ادغام و ایجاد منطقه‌ی جذاب پارک High Line (ب) سه خوشه‌ی کشف شده به صورت جداگانه و در مجاورت یکدیگر، پیش از ادغام



تصویر ۵- در قسمت (الف) موزه‌ی پایتخت هنر با طول حدود ۲۰۰ متر به عنوان منطقه‌ی جذاب کوچک و متراکم و قسمت (ب) فرودگاه J.F.Kennedy با طول حدود ۴۵۰۰ متر به عنوان منطقه‌ی جذاب بزرگ و کم تراکم را نشان می‌دهد.

۵-۳- ایجاد اطلاعات معنایی

برچسب‌های برتر هر منطقه‌ی جذاب به شکل ابر برچسب با استفاده از کتابخانه‌ی wordcloud در پایتون ایجادشده است که نمونه‌ی آن در تصویر ۷ نشان داده شده است. همچنین نام و دسته‌بندی مربوط به هر منطقه‌ی جذاب، با استفاده از برچسب‌های متنی و حدود ۱۷ هزار نقطه‌ی جذاب پایگاه داده‌ی GeoName در محدوده‌ی

در مرحله‌ی پیش‌پردازش، تعداد ۱۰۸۲۱ رکورد داده‌ی تکراری، ۹۰۴۰۵ رکورد داده با دقت تعیین موقعیت کم، ۲۳۷۵ رکورد داده با برچسب‌های مربوط به مراسم شخصی و ۱۸۴۰۶۳ رکورد داده که در فاصله‌ی زمانی کمتر از یک دقیقه توسط یک کاربر اخذشده بودند، حذف شدند. به این ترتیب ۳۳۴۸۴۶ رکورد داده پس از پیش‌پردازش باقی ماندند که توسط ۱۴۳۴۶ کاربر یکتا اخذشده‌اند.

۲-۳- خوشه‌بندی مکانی

ابتدا مختصات مکانی داده‌ها که به صورت طول و عرض جغرافیایی در سیستم مختصات WGS84 قرار دارند، با استفاده از سیستم تصویر UTM به صورت متریک درآورده می‌شوند تا عملیات خوشه‌بندی و مراحل بعد انجام پذیر شوند. سپس با استفاده از کتابخانه‌ی hdbscan در پایتون عملیات خوشه‌بندی انجام می‌شود. با سعی و خطای انجام شده، پارامتر اجباری "min_cluster_size" روی ۴۵۰ و پارامتر اختیاری "min_samples" روی ۱۰۰ تنظیم شده است. در نتیجه‌ی عملیات خوشه‌بندی HDBSCAN تعداد ۱۱۶ خوشه کشف شده است.

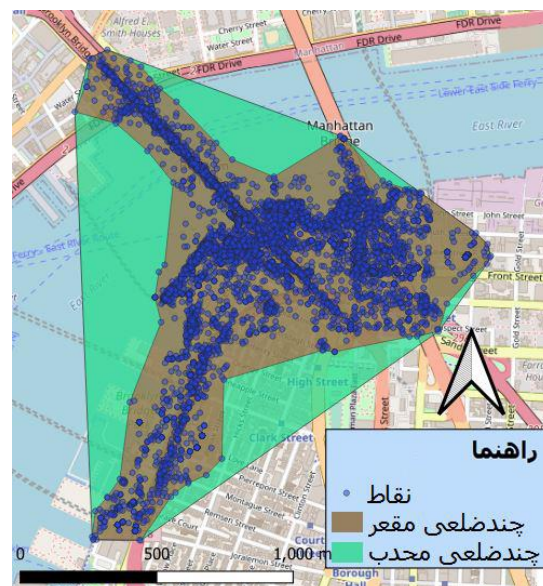
۳-۳- ادغام خوشه‌ها

با بررسی خوشه‌بندی انجام شده، خوشه‌هایی که در مجاورت یکدیگر قرار دارند و شباهت بردار برچسب‌های وزن دار شده‌ی آن‌ها، بیش از ۳۵ درصد بوده، در یکدیگر ادغام شده‌اند که در نتیجه تعداد ۱۰ خوشه در خوشه‌های مجاورشان ادغام شدند و تعداد کل خوشه‌ها به ۱۰۶ عدد تقلیل یافت. در تصویر ۴ نمونه‌ای از این فرآیند برای پارک High Line در نیویورک، قبل و بعد از ادغام نشان داده شده‌اند.

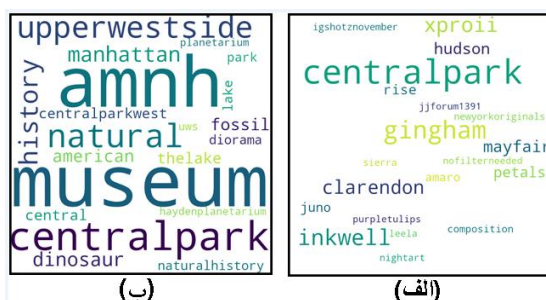
۴-۳- ایجاد مرز مناطق جذاب

با استفاده از کتابخانه‌ی alphashape و geopandas در پایتون، چندضلعی تعیین کننده‌ی مرز مناطق جذاب ساخته شده‌اند. مقدار پارامتر alpha که تعیین کننده‌ی میزان تقعر شکل ایجادشده است، از طریق تابع optimizealpha در کتابخانه‌ی alphashape به صورت بهینه برای هر خوشه محاسبه شده است. نمونه‌ای از دو

شهر نیویورک که حدود ۲۲۰۰ عدد از آن‌ها در محدوده‌ی چندضلعی‌های استخراج‌شده قرار داشتند، استخراج شدند. نمونه‌ای از نام و دسته‌بندی انتخاب‌شده در جدول ۲ آورده شده است.



تصویر ۶- مقایسه‌ی فضای خالی ایجادشده در چندضلعی برازش داده‌شده با روش سطح محدب در مقایسه با چندضلعی با روش سطح مقعر در منطقه‌ی جذاب پل بروکلین



تصویر ۷- ابر کلمات ایجادشده برای (الف) منطقه‌ی جذاب پارک Central و (ب) موزه‌ی تاریخ طبیعی آمریکا

۳-۶- تحلیل مناطق جذاب

۳-۶-۱- تحلیل احساسات کاربران

بررسی عقاید و احساسات کاربران نسبت به منطقه‌ی جذاب با استفاده از نظرات کاربران و کتابخانه‌ی TextBlob در پایتون انجام شد. این کتابخانه مبتنی بر کتابخانه‌ی دیگری به نام pattern است و با استفاده از برچسب‌گذاری اجزای جملات و استفاده از مجموعه داده‌ای از کلمات، تحلیل عقاید را انجام می‌دهد و برای خروجی عددی بین ۱- تا ۱+ را ارائه می‌دهد که ۱- به معنای نظر کاملاً منفی و

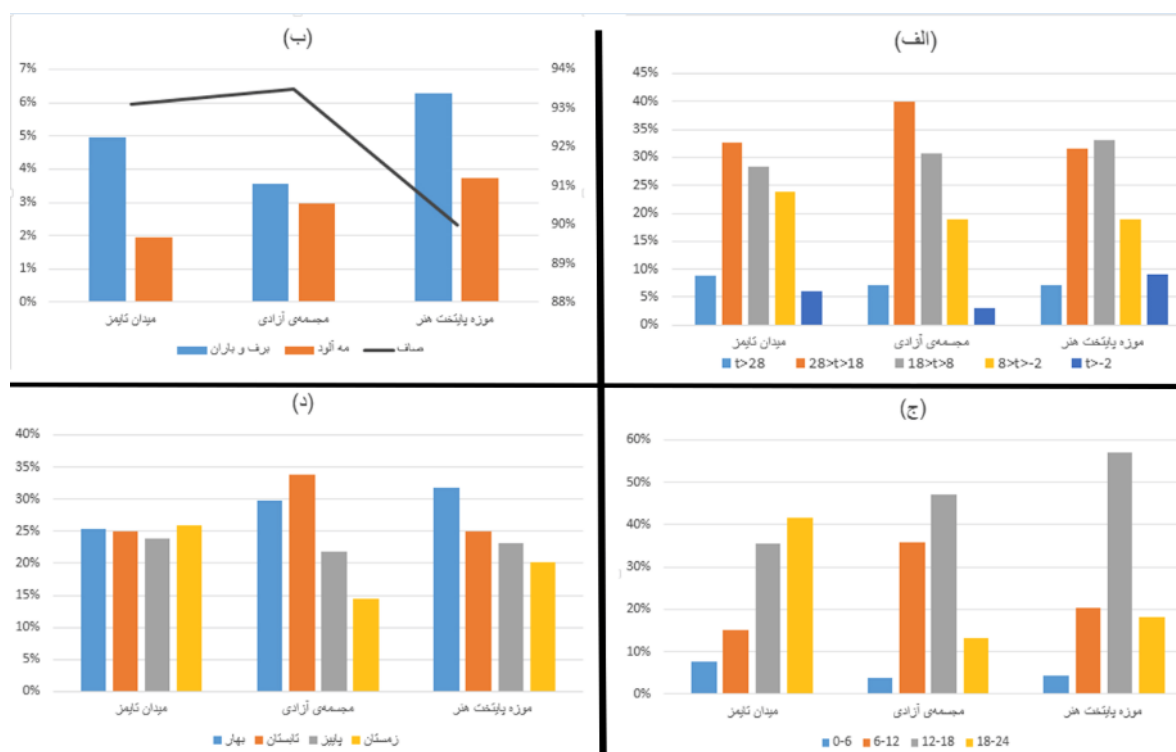
۱+ نظر کاملاً مثبت و صفر خنثی تلقی می‌شود [۲۷]. نظرات خنثی با مقادیر حدود صفر حذف شدند و از کل نظرات هر منطقه میانگین گرفته شده است. بیشترین مقدار احساسات مثبت برای منطقه‌ی Rockefeller Center با مقدار ۰/۷۶ و کمترین آن برای استادیوم Jack Coffey با مقدار ۰/۳۶ بوده است. در جدول ۲ تعدادی از مناطق جذاب به همراه تحلیل احساسات انجام‌شده در مقیاس ۵ ستاره‌ای آورده شده است.

۳-۶-۲- کاربران گردشگر و محلی

با استفاده از روش ذکرشده، میزان محتوای مربوط به گردشگران و شهروندان محلی، در هر منطقه‌ی جذاب بررسی شد. بالاترین مقدار برای جزیره‌ی Ellis با ۶۹ درصد گردشگر و کمترین آن محله‌ی Prospect Heights با کمتر از یک درصد گردشگر بوده است. در جدول ۲ نمونه‌ای از مناطق جذاب کشف‌شده به همراه درصد کاربرانی را که گردشگر بوده‌اند نشان داده شده است.

جدول ۲- تعدادی از مناطق جذاب مشهور در شهر نیویورک به همراه نام و دسته‌بندی استخراج‌شده، درصد گردشگران و احساسات تحلیل‌شده نشان داده شده است.

نام	دسته‌بندی	احساسات	گردشگران
Brooklyn Bridge	پل	4.2	35%
Times Square	میدان	4.22	43%
Oculus	ایستگاه قطار	4.48	44%
High Line	پارک	4.18	33%
Rockefeller Center	ساختمان	4.53	49%
Empire State Building	ساختمان	4.27	46%
Union Square	میدان	4.3	13%
Statue of Liberty	بنای یادبود	4.3	68%



تصویر ۸- تعدادی از مناطق جذاب در شرایط بافت مختلف مقایسه شده‌اند. (الف) در بازه‌های مختلف دمای هوا. (ب) در شرایط آب و هوایی مختلف (ج) در بازه‌های زمانی در طول روز. (د) در فصل‌های سال مقایسه شده‌اند.

۳-۶-۳- تحلیل شرایط بافت مختلف

هر منطقه‌ای جذاب کشف‌شده، در شرایط مختلفی بافتی بررسی می‌شود که نمونه‌ای از آن در تصویر ۸ نشان داده شده است. در بررسی نمونه‌های انتخابی، موزه پایتخت هنر در دماهای بسیار سرد یا بسیار گرم و همچنین در شرایط بارش‌های جوی، بازدیدکنندگان بیشتری نسب به سایر مناطق منتخب داشته است که احتمالاً به بسته بودن محیط آن مرتبط است. همچنین میدان تایمز که در مرکز بخش منتهن واقع شده است، دارای بازدیدکننده‌ی تقریباً یکسانی در فصل‌های مختلف سال بوده و برخلاف سایر مناطق منتخب در بازه‌ی پایانی شبانه‌روز بازدیدکننده‌ی بیشتری داشته است و مجسمه‌ی آزادی در فصل‌های سردتر سال و در دماهای پایین، کمتر موردتوجه قرار گرفته است که احتمالاً به دلیل قرار گرفتن در محیط دریایی و مخاطبان گردشگر آن است.

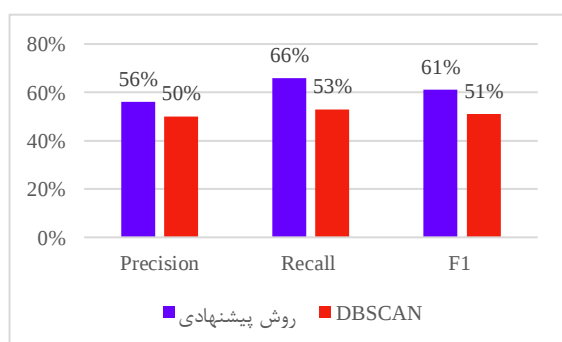
۳-۷-۳- ارزیابی

از طریق وبگاه TripAdvisor که حاوی اطلاعات تولیدشده توسط کاربران است، تعداد ۲۰۰ جاذبه‌ی برتر بر اساس تعداد نظرات، تصاویر و رتبه‌های کاربران در شهر

نیویورک اخذ شدند تا برای مقایسه با مناطق جذاب کشف‌شده در این تحقیق مورداستفاده قرار گیرند. اطلاعات اخذشده شامل موقعیت، نام و بازخورد کاربران در مورد جاذبه‌های شهر نیویورک است [۲۸].

۳-۷-۱- ارزیابی مکانی

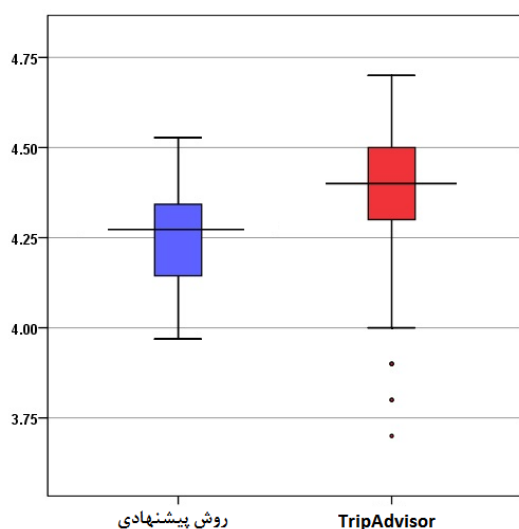
جهت مقایسه کمی، روش DBSCAN ساده با استفاده از بسته‌ی Scikitlearn در پایتون پیاده‌سازی شده و پارامتر شعاع جستجو ۵۰ متر و پارامتر آستانه‌ی تراکم ۲۴۰ در نظر گرفته شده که در نتیجه‌ی آن ۱۰۹ منطقه‌ی جذاب کشف شده است. نتایج حاصل از دو روش در تصویر ۹ نمایش داده شده است.



تصویر ۹- مقایسه‌ی شاخص‌های ارزیابی روش پیشنهادی و DBSCAN

۳-۷-۳- ارزیابی تحلیل احساسات

کاربران در TripAdvisor در کنار متن نظرات خود، رتبه‌ای بین ۱ تا ۵ ستاره به جاذبه‌ی موردنظر می‌دهند که ۱ ستاره به معنی خیلی بد، ۲ ستاره به معنی ضعیف، ۳ ستاره به معنی متوسط، ۴ ستاره به معنی خوب و ۵ ستاره به معنی عالی است. با استفاده از تعداد نظرات در هر حالت و گرفتن میانگین وزن‌دار از آن‌ها، تعداد ستاره مربوط به هر جاذبه در TripAdvisor به صورت عددی اعشاری بین ۱ تا ۵ به دست آمده است. همچنین با تحلیل احساسات نظرات کاربران در مورد مناطق جذاب کشف‌شده که عددی بین +۱ تا -۱ است سپس با جمع آن با عدد ۱/۵ و ضرب آن در عدد ۲ به حالت ۱ تا ۵ ستاره‌ای تبدیل و با داده‌های TripAdvisor هم‌مقیاس شده‌اند.



تصویر ۱۰- نمودار جعبه‌ای جهت مقایسه‌ی رتبه‌ی داده‌شده در TripAdvisor و رتبه‌ی حاصل از تحلیل احساسات نظرات کاربران ارائه شده است. محور قائم رتبه داده شده در مقیاس ۵ ستاره‌ای را نشان می‌دهد

در تصویر ۱۰ تحلیل احساسات انجام شده در تحقیق برای مناطق جذابی که با جاذبه‌های TripAdvisor هم‌پوشانی داشته‌اند، با ستاره‌های بازخورد کاربران در مورد جاذبه‌های TripAdvisor به صورت نمودار جعبه‌ای مقایسه شده است. همچنین جهت مقایسه مقدار RMSE^۴ بین آن‌ها محاسبه شده و مقدار ۰/۱۴۱ به دست آمده است.

طبق تصویر ۸ حدود ۵۶ درصد از مناطق جذاب کشف‌شده شامل جاذبه‌های TripAdvisor هستند و همچنین ۶۶ درصد از کل جاذبه‌های TripAdvisor در مناطق جذاب کشف‌شده قرار گرفتند که نشان می‌دهد تعداد قابل توجهی مناطق جذاب کشف شده است که با وجود تعامل زیاد کاربران با آن مناطق، از منظر TripAdvisor جزو مناطق جذاب برتر قرار نداشته‌اند که نمونه‌های آن را می‌توان باغ وحش Queens، باغ وحش Bronx، فرودگاه J.F.Kennedy، پارک Brox و ورزشگاه Citi Field نام برد و دلیل آن را می‌توان تفاوت دیدگاه نسبت به جذابیت مکان‌های برتر دانست، زیرا مناطقی که کاربران تعامل بیشتری داشته‌اند در این تحقیق جذاب محسوب می‌شود ولی از دیدگاه TripAdvisor جذابیت مکان‌ها بر اساس رتبه‌ها، نظرات و تصاویر کاربران محاسبه شده است. همچنین عمده‌ی جاذبه‌های معرفی‌شده در TripAdvisor در بخش منهتن^۱ شهر نیویورک قرار گرفته‌اند و تعداد کمی در سایر بخش‌های شهر هستند، این در حالی است که تعداد قابل توجهی از مناطق جذاب کشف‌شده در تحقیق در بخش‌های بروکلین^۲ و برانکس^۳ قرار دارند و می‌توان این گونه تفسیر کرد که روش پیشنهادی در این تحقیق توانایی کشف مناطق جذاب محلی و با شهرت کمتر را نیز دارد. همچنین مطابق تصویر ۹ هر سه پارامتر ارزیابی در روش پیشنهادی، نسبت به روش DBSCAN ساده، مقادیر بهتری را کسب کرده است.

۳-۷-۲- ارزیابی نام‌های انتخاب‌شده

برای نشان دادن کارایی روش ارائه‌شده جهت نام‌گذاری مناطق جذاب، بدیهی است که مناطق جذابی که از نظر مکانی با جاذبه‌های TripAdvisor پوششی نداشته‌اند مورد بررسی قرار نمی‌گیرند و تنها نام مناطقی که با جاذبه‌های TripAdvisor هم‌پوشانی داشته‌اند باید بررسی شود؛ بنابراین همان‌طور که در بخش پیشین بیان شد، از بین کل ۱۰۶ منطقه‌ی جذاب کشف‌شده ۵۶ درصد یا ۶۰ عدد از آن‌ها با جاذبه‌های TripAdvisor هم‌پوشانی داشتند و بایستی نام انتخاب‌شده برای آن‌ها بررسی شود. از میان این ۶۰ منطقه‌ی جذاب، برای ۴۶ عدد یا به عبارتی ۷۶ درصد از آن‌ها اسم مشابهی توسط روش ارائه‌شده، انتخاب شده است.

^۱ Manhattan

^۲ Brooklyn

^۳ Bronx

^۴ Root-Mean-Square Error

۵- نتیجه گیری و پیشنهادها

در این تحقیق با استفاده از تصاویر دارای برچسب مکانی فلیکر به کشف و تحلیل مناطق جذاب شهر نیویورک پرداخته شد. برای این منظور با حذف نوفه و افزودگی داده‌ها در مرحله‌ی پیش‌پردازش، با استفاده از روش HDBSCAN داده‌ها خوشه‌بندی شدند و خوشه‌های مجاور که از نظر برچسب‌های متنی مشابه بودند ادغام شدند و ۱۰۶ منطقه‌ی جذاب شناسایی شد. سپس با استفاده از روش α -shape سطح مقعرى به نقاط برازش شد و اطلاعات معنایی مانند برچسب متمایز، نام و دسته‌بندی برای مناطق جذاب انتخاب شد. سپس مناطق جذاب بر اساس نوع کاربران بازدیدکننده (شهروند و گردشگر)، احساسات کاربران در نظرات، میزان بازدید در شرایط بافت مختلف تحلیل شدند.

ارزیابی نتایج نشان داد روش پیشنهادی از نظر مکانی ۶۶ درصد از جاذبه‌های TripAdvisor را شامل شده و در کنار جاذبه‌های مشهور در بخش منهن توانسته جاذبه‌های محبوب و محلی و با شهرت کمتر را هم کشف کند و نتایج بهتری نسبت به DBSCAN ساده ارائه کند. بررسی‌ها همچنین نشان داد مناطق جذاب با اندازه‌های متفاوتی از چند صد متر تا چند کیلومتر کشف شده که نمونه‌ی آن در قسمت قبل آورده شده است و نشان می‌دهد مسئله‌ی تفاوت در تراکم تا حدودی رفع گردیده است. این در حالی است که در روش‌هایی مثل DBSCAN در صورتی که

پارامتر شعاع جستجو کوچک انتخاب شود، خوشه‌های بزرگ و کم تراکم کشف نمی‌شوند یا با ابعاد کوچکی کشف می‌شوند و اگر بزرگ انتخاب شود ممکن است منطقه‌ای به بزرگی بخش منهن نیویورک، یک خوشه در نظر گرفته شود. همچنین ارزیابی نشان داد نام‌گذاری در روش پیشنهادی در ۷۶ درصد از مناطقی که همپوشانی داشتند نتایج مشابهی ارائه کرده است و مقادیر تحلیل عقاید نیز در محدوده‌ی مشابهی به دست آمده است.

برای تحقیقات آینده استفاده از سایر منابع داده‌ی دارای برچسب مکانی و استفاده ترکیبی از آن‌ها برای حل مشکل کمبود داده‌ها در شهرهای کوچک پیشنهاد می‌شود. همچنین استفاده از سایر روش‌های خوشه‌بندی پیشنهاد می‌شود، به گونه‌ای که در کنار برچسب مکانی از سایر اطلاعات موجود مانند محتوای تصویری، برچسب‌های متنی و مختصات ارتفاعی هم استفاده شود. همچنین استفاده از ویژگی ارتفاعی و خوشه‌بندی سه‌بعدی نقاط نیز قابل بررسی است. در راستای بهبود اطلاعات معنایی، به جای استفاده از منبع داده‌ی اضافی، استفاده از محتوای تصویری جهت کشف دسته‌بندی و نوع منطقه‌ی جذاب پیشنهاد می‌گردد. استفاده از روش‌های مبتنی بر موارد و مبتنی بر قواعد به صورت ترکیبی، به عنوان روشی نوین در تحلیل عقاید قابل بررسی است. در نهایت استفاده از نتایج به دست آمده برای ارائه‌ی سیستم‌های توصیه‌گر مناطق جذاب پیشنهاد می‌شود.

مراجع

- [1] Batty, M. (2013). "Big Data, Smart Cities and City Planning." *Dialogues in human geography*, vol. 3, no. 3, pp. 274-279 .
- [2] Giglio, S., Bertacchini, F., Bilotta, E. and Pantano, P. (2019). "Using Social Media to Identify Tourism Attractiveness in Six Italian Cities." *Tourism Management*, vol. 72, pp. 306-312 .
- [3] Tjepkema, L. "Top 5 Social Media Predictions for 2019." Emarsys. <https://emarsys.com/learn/blog/top-5-social-media-predictions-2019/> (accessed)
- [4] AlBanna, B., Sakr, M., Moussa, S. and Moawad, I. (2016). "Interest Aware Location-Based Recommender System Using Geo-Tagged Social Media." *ISPRS International Journal of Geo-Information*, vol. 5, no. 12, pp. 245 .
- [5] Kuo, C.-L., Chan, T.-C., Fan, I. and Zipf, A. (2018). "Efficient Method for Poi/Roi Discovery Using Flickr Geotagged Photos." *ISPRS International Journal of Geo-Information*, vol. 7, no. 3, pp. 121 .
- [6] Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W. and Prasad, S. (2015). "Extracting and Understanding Urban Areas of Interest Using Geotagged Photos." *Computers, Environment and Urban Systems*, vol. 54, pp. 240-254 .
- [7] Zheng, Y.-T., Zha, Z.-J. and Chua, T.-S. (2012). "Mining Travel Patterns from Geotagged Photos." *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, pp. 1-18 .

- [8] García-Palomares, J. C., Gutiérrez, J. and Mínguez, C. (2015). "Identification of Tourist Hot Spots Based on Social Networks: A Comparative Analysis of European Metropolises Using Photo-Sharing Services and Gis." *Applied Geography*, vol. 63, pp. 408-417 .
- [9] Cheng, A.-J., Chen, Y.-Y., Huang, Y.-T., Hsu, W. H. and Liao, H.-Y. M. (2011). "Personalized Travel Recommendation by Mining People Attributes from Community-Contributed Photos," presented at the Proceedings of the 19th ACM international conference on Multimedia.
- [10] Peng, X. and Huang, Z. (2017). "A Novel Popular Tourist Attraction Discovering Approach Based on Geo-Tagged Social Media Big Data." *ISPRS International Journal of Geo-Information*, vol. 6, no. 7, pp. 216 .
- [11] Lee, J. Y. and Tsou, M.-H. (2018). "Mapping Spatiotemporal Tourist Behaviors and Hotspots through Location-Based Photo-Sharing Service (Flickr) Data," presented at the LBS 2018: 14th International Conference on Location Based Services.
- [12] Paldino, S., Bojic, I., Sobolevsky, S., Ratti, C. and González, M. C. (2015). "Urban Magnetism through the Lens of Geo-Tagged Photography." *EPJ Data Science*, vol. 1, no. 4, pp. 1-17 .
- [13] Mou, N., Yuan, R., Yang, T., Zhang, H., Tang, J. J. and Makkonen, T. (2020). "Exploring Spatio-Temporal Changes of City Inbound Tourism Flow: The Case of Shanghai, China." *Tourism Management*, vol. 76, pp. 103955 .
- [14] Spyrou, E., Korakakis, M., Charalampidis, V., Psallas, A. and Mylonas, P. (2017). "A Geo-Clustering Approach for the Detection of Areas-of-Interest and Their Underlying Semantics." *Algorithms*, vol. 10, no. 1, pp. 35 .
- [15] KhoshAmooz, G. and Taleai, M. (2017). "Identification and Ranking Regions of Interest Based on Volunteer-Employed Photography and Geo-Tagged Photos." *Journal of Geomatics Science and Technology*, vol. 7, no. 1, pp. 15-28. (in Persian)
- [16] Encalada, L., Boavida-Portugal, I., Cardoso Ferreira, C. and Rocha, J. (2017). "Identifying Tourist Places of Interest Based on Digital Imprints: Towards a Sustainable Smart City." *Sustainability*, vol. 9, no. 12, pp. 2317 .
- [17] Omrani khiyabaniyan, S., Alesheikh, A. and Abbasi, O. (2019). "Developing a Recommendation Framework for Tourist by Mining Geo-Tag Photos (Case Study Tehran District 6)." *Journal of Geomatics Science and Technology*, vol. 9, pp. 31-42. (in Persian)
- [18] Majid, A., Chen, L., Mirza, H. T., Hussain, I. and Chen, G. (2015). "A System for Mining Interesting Tourist Locations and Travel Sequences from Public Geo-Tagged Photos." *Data & Knowledge Engineering*, vol. 95, pp. 66-86 .
- [19] Flickr. "Flickr Photos Search." <https://www.flickr.com/services/api/flickr.photos.search.htm> (accessed)
- [20] Campello, R. J., Moulavi, D. and Sander, J. (2013). "Density-Based Clustering Based on Hierarchical Density Estimates", presented at the Pacific-Asia conference on knowledge discovery and data mining.
- [21] Edelsbrunner, H., Kirkpatrick, D. and Seidel, R. (1983). "On the Shape of a Set of Points in the Plane." *IEEE Transactions on information theory*, vol. 29, no. 4, pp. 559-551 .
- [22] Liu, M. and Yang, J. (2012). "An Improvement of Tfidf Weighting in Text Categorization." *International proceedings of computer science and information technology*, vol. 47, pp. 44-47 .
- [23] Hopfgartner, F., (2015). "Smart Information Systems :Computational Intelligence for Real-Life Applications." Springer.
- [24] Berka, P. (2020). "Sentiment Analysis Using Rule-Based and Case-Based Reasoning." *Journal of Intelligent Information Systems*, vol. 55, pp. 1-16 .
- [25] Olson, D. L. and Delen, D., (2008) .(Advanced Data Mining Techniques." Springer Science & Business Media.
- [26] Wikipedia. "New York City." https://en.wikipedia.org/wiki/New_York_City (accessed)
- [27] Loria, S. (2018). "Textblob Documentation." Release 0.15, vol. 2 ,
- [28] Tripadvisor. "New York City Attractions." https://www.tripadvisor.com/Attractions-g60763-Activities-a_allAttractions.true-New_York_City_New_York.html (accessed)