نشریه علمی علوم و فنون نقشه برداری، دوره دهم، شماره ۲، آذر ماه ۱۳۹۹ - مقاله پژوهشی

مکان‌مرجع‌سازی منابع محل‌مبنای نیمه‌ساختاریافته در وب با استفاده از یادگیری ماشین

امیدرضا عباسی*[۱]، علی اصغر آل شیخ[۲]

[۱] دانشجوی دکتری سیستم‌های اطلاعات مکانی - دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه نصیرالدین طوسی
oabbasi@mail.kntu.ac.ir

[۲] استاد دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه نصیرالدین طوسی
alesheikh@kntu.ac.ir

(تاریخ دریافت اسفند ۱۳۹۸، تاریخ تصویب آبان ۱۳۹۹)

چکیده

در سال‌های اخیر محتوای منتشرشده بر روی وب به طور چشمگیری افزایش یافته است. بخش عمده‌ای از این اطلاعات به‌صورت نیمه‌ساختاریافته در اختیار عموم قرار دارند. علاوه بر این، حجم عظیمی از اطلاعات مرتبط با محل هستند. این گونه اطلاعات به یک مکان بر روی زمین اشاره دارند، اما دارای مختصات صریح آن محل نیستند. در این مقاله به مکان‌مرجع‌سازی منابع نیمه‌ساختاریافته در وب با استفاده از یادگیری ماشین پرداخته شده است. مزیت این روش عدم نیاز به استفاده از روش‌های پیچیده متن‌کاوی جهت مکان‌مرجع‌سازی است. بدین منظور، از آگهی‌های تبلیغاتی تارنمای دیوار مرتبط با املاک و مستغلات در شهر تهران استفاده شده است. به منظور جمع‌آوری داده از روش خزیدن در وب بهره برده شده است. همچنین، جهت دستیابی به هدف تحقیق، الگوریتم جنگل‌های تصادفی به عنوان یک روش یادگیری ماشین مناسب بکار گرفته شده است. نتایج تحقیق نشان می‌دهد که با استفاده از این روش، آگهی‌های تبلیغاتی تارنمای دیوار با دقت مناطق تهران قابل مکان‌مرجع‌سازی است. به طور کمّی، دقت بدست آمده در این تحقیق حدود ۶ کیلومتر در راستای طول جغرافیایی و حدود ۲ کیلومتر در راستای عرض جغرافیایی است. همچنین، نتایج این تحقیق نشان می‌دهد که متغیر قیمت ملک نسبت به دیگر متغیرها از اهمیت و تاثیر بیشتری در تعیین مکان آگهی برخوردار است. علاوه بر این، از نتایج این تحقیق می‌توان نتیجه گرفت که قیمت املاک در شهر تهران در راستای شمالی - جنوبی دارای الگوی مکانی بیشتری نسبت به راستای شرقی - غربی هستند.

واژگان کلیدی: مکان‌مرجع‌سازی، اطلاعات محل‌مبنا، منابع تحت وب، جنگل‌های تصادفی

* نویسنده رابط

۱- مقدمه

با گسترش روزافزون میزان محتوای متنی در حال انتشار بر بستر وب نظیر وب‌نوشت‌ها، شبکه‌های اجتماعی و سکوهای به‌اشتراک‌گذاری محتواهای چندرسانه‌ای، نیاز به روش‌هایی جهت درک خودکار آن‌ها احساس می‌شود [۱]. بخش بزرگی از این محتوا به صورت اطلاعات متنی نیمه‌ساختار یافته منتشر می‌شود. اطلاعات نیمه-ساختاریافته عمدتاً به‌صورت برچسب‌گذاری شده یا به-صورت مجموعه‌ای از خصوصیات و مقادیر به اشتراک-گذاری می‌شوند. این روش در شبکه‌های اجتماعی بسیاری نظیر اینستاگرام مورد استفاده است. به اطلاعات مرتبط به مکان موجود در متون، گفتار و محاورات اطلاعات محل-مبنا گفته می‌شود [۲]. در این اطلاعات مختصات جغرافیایی به‌طور صریح یا دقیق ذکر نشده، اما به گونه‌ای به یک مکان اشاره دارند. کاربران در توصیف محل معمولا (به صورت ناخودآگاه) یک سیستم مختصات کیفی ایجاد و از آن برای تعیین محل موردنظر خود استفاده می‌کنند.

مکان مرجع‌سازی به فرایندی گفته می‌شود که در آن مختصات جغرافیایی یک منبع پیش‌بینی و به آن منبع تخصیص داده می‌شود. امروزه بسیاری از شبکه‌های اجتماعی به کاربر این امکان را داده‌اند تا با استفاده از GPS تلفن‌های همراه خود یا با فراهم کردن یک نقشه مکان مورد نظر خود را به اشتراک بگذارند. با این حال، بیشتر محتوای وب فاقد مختصات صریح هستند و این موضوع نشان‌دهنده نیاز به توسعه روشی جهت مکان-مرجع‌سازی خودکار محتوای وب است. یکی از اولین مراحل جهت آماده‌سازی داده‌های محل‌مبنا برای انجام تحلیل‌های مکانی مکان‌مرجع‌سازی آن‌ها است. در واقع، به‌منظور توانمندسازی رایانه‌ها در درک این اطلاعات لازم است محل‌ها در فضا مکان‌مرجع شوند. علاوه بر این، مکان‌مرجع‌سازی محتوای وب کاربردهای فراوانی در خدمات مکان‌مبنا [۳]، بازیابی اطلاعات مکانی [۴] و تعامل کاربر با رایانه [۵] دارد.

در این تحقیق به مکان‌مرجع‌سازی آگهی‌های فروش و اجاره املاک در تارنمای دیوار با استفاده از الگوریتم یادگیری ماشین جنگل‌های تصادفی پرداخته شده است. آگهی‌های مورد نظر به عنوان یک ارجاع به محل در نظر گرفته می‌شود. در واقع، هر آگهی به یک ملک اشاره دارد و هر ملک دارای مختصاتی است. تا به حال از روش‌های مختلفی جهت مکان‌مرجع‌سازی اطلاعات بدون ساختار متنی استفاده شده است. با این حال، تفاوت اطلاعات ساختاریافته با اطلاعات متنی بدون ساختار این است که در این نوع اطلاعات می‌توان بدون نیاز به متن‌کاوی از الگوریتم‌های یادگیری ماشین استفاده نمود [۶]. در واقع، با استفاده از مجموعه ویژگی‌های یک محل، شامل ویژگی-های مکانی و غیرمکانی، می‌توان با دقت مناسب به مختصات محل دست یافت. این در حالی است که روش-های بکارگرفته شده برای مکان‌مرجع‌سازی محل از اطلاعات متنی بدون ساختار عمدتاً برای عوارض جغرافیایی با وسعت زیاد نظیر شهرها و کشورها و کوه‌ها و رودها طراحی شده‌اند [۷]. در این تحقیق سعی بر آن است تا املاک آگهی شده در تارنمای دیوار به‌عنوان عوارض جغرافیایی مکان‌مرجع شوند. علاوه بر این، در این روش نیازی به پردازش‌های نسبتاً سنگین نظیر پردازش زبان طبیعی و متن‌کاوی نیست.

در ادامه این مقاله، در بخش ۲ به مرور تحقیقات پیشین خواهیم پرداخت. در بخش ۳ مبانی نظری تحقیق شامل مفاهیم مرتبط با مکان‌مرجع‌سازی، اطلاعات محل-مبنا و الگوریتم یادگیری ماشین جنگل‌های تصادفی پرداخته می‌شود. بخش ۴ داده‌های تحقیق، نحوه پیاده-سازی و نتایج تحقیق را ارائه می‌دهد. در نهایت در بخش ۵ به نتیجه‌گیری و بیان محدودیت‌ها و ارائه پیشنهادهای آتی خواهیم پرداخت.

۲- تحقیقات پیشین

در این بخش به مرور تحقیقات انجام‌شده در حوزه زمین‌مرجع‌سازی محتوای وب و شبکه‌های اجتماعی می-پردازیم. یکی از ایده‌های موجود جهت زمین‌مرجع‌سازی منابع متنی، یافتن جای‌نام‌ها[۱] در متن است. جای‌نام‌ها ممکن است به صورت یک آدرس باشند یا اشاره مستقیم به نام منطقه موردنظر داشته باشند. در هر صورت، پس از یافتن جای‌نام‌ها در متن، موقعیت آن‌ها با استفاده از مداخل مکانی[۲] تعیین می‌شود. مداخل مکانی دربردارنده جای‌نام‌های رسمی نظیر نام شهرها، استان‌ها و عوارض

۱ Placenames
۲ Gazetter

برجسته و موقعیت‌های متناظر آن‌ها هستند. سرویس GeoNames یکی از بزرگ‌ترین مداخل موجود است و بیش از ۱۱ میلیون مدخل دارد.

به عنوان مثال، بوسکالدی و روسو [۸] با استفاده از مدخل Wikipedia-World به زمین‌مرجع‌سازی مداخل هستان‌شناسی[۱] WordNet پرداخته‌اند. WordNet یک پایگاه داده واژگان انگلیسی است که در آن مفاهیم هم‌معنا با استفاده از روابط مفهومی مرتبط شده‌اند.

یکی از معایب این روش آن است که جای‌نام‌ها در برخی موارد به عنوان اسم خاص برای یک محل به کار نمی‌روند. به عنوان مثال، واژه «جمهوری» می‌تواند به «خیابان جمهوری» یا به شیوه حکومت در یک کشور اشاره داشته باشد. بنابراین، این روش نیازمند فرایند ابهام‌زدایی[۲] است. همچنین، محل‌های مختلف بر روی زمین ممکن است جای‌نام‌های یکسانی داشته باشند. برای مثال، «آستارا» نام شهرستانی در شمال ایران و نام شهرستانی در جنوب کشور جمهوری آذربایجان است. از این رو، برای کشف مکان واقعی این جای‌نام‌ها ابتدا لازم است مشخص شود که به کدام منطقه اشاره دارد. به این فرایند جداسازی جای‌نام‌ها[۳] گفته می‌شود. یکی دیگر از مشکلاتی که زمین-مرجع‌سازی از طریق متون و مداخل با آن مواجه است، استفاده کاربران از جای‌نام‌های غیررسمی و روزمره است. برای رفع این مشکل معمولاً به جمع‌آوری اطلاعات در مورد این نام‌ها از منابع وب و شبکه‌های اجتماعی می‌پردازند.

فینک و همکاران [۹] با هدف تعیین موقعیت نویسندگان وب‌نوشت‌ها[۴] به تحلیل متن نوشته‌های آن‌ها پرداخته‌اند. در این پژوهش ابتدا جای‌نام‌های موجود در نوشته‌ها استخراج و با استفاده از مدخل GeoNames مطابقت داده می‌شوند. همچنین، جای‌نام‌های دارای ابهام از محاسبات کنار گذاشته شده‌اند و مرحله ابهام زدایی حذف شده است. سپس با استفاده از خوشه‌بندی تمام موقعیت‌ها و انتخاب یک مرکز خوشه، موقعیت نویسنده مشخص می‌شود. دقت مکان‌مرجع‌سازی در این پژوهش در سطح کشور ۹۷ درصد و در سطح شهرستان ۶۵ درصد بوده است.

چن و همکاران [۱۰] با استفاده از گراف‌های محلی روش جدیدی ارائه کرده‌اند که در آن می‌توان محل‌ها را بدون توجه به رسمی بودن نام آن‌ها، یعنی وجود نام آن‌ها در مداخل مکانی، زمین مرجع نمود. در این پژوهش ابتدا یک گراف محلی از اطلاعات توصیفی ایجاد می‌شود. این گراف به عنوان پایگاه دانش برای محل‌ها و روابط مکانی بین آن‌ها در نظر گرفته می‌شود. در مرحله بعد، ابتدا جای‌نام‌هایی که در مداخل مکانی وجود دارند زمین مرجع می‌شوند. سپس، مکان جای‌نام‌هایی که در مداخل وجود ندارند، با استفاده از روابط مکانی موجود در گراف به طور تقریبی به‌دست می‌آیند. بیشینه دقت زمین‌مرجع‌سازی در این تحقیق به ۹۰ درصد می‌رسد. با این وجود، این روش در مورد محل‌هایی کاربرد دارد که روابط مکانی آن‌ها با محل‌های موجود در مداخل مکانی صراحتاً ذکر شده باشد.

گروور و همکاران [۱۱] با استفاده از تجزیه‌کننده مکانی[۵] Edinburgh به زمین‌مرجع‌سازی منابع تاریخی پرداخته‌اند. در این پژوهش، ابتدا جملات تقسیم به قطعات کوچک‌تر شده و عملیات تکه‌تکه‌سازی[۶] بر روی آن‌ها انجام می‌شود. سپس اجزای کلام[۷] در این قطعات شناسایی و اسامی خاص از آن‌ها جدا می‌شوند. در نهایت، بمنظور ابهام‌زدایی و تخصیص مختصات به واژه از یک مدخل استفاده می‌شود.

علاوه بر پژوهش‌هایی که در آن‌ها از روش‌های متن-کاوی جهت تعیین موقعیت منبع استفاده می‌شود، روش-هایی وجود دارند که با استفاده از اطلاعات مرتبط دیگر به این مهم می‌پردازند.

کرندال و همکاران [۱۲] با استفاده از روش‌های خوشه‌بندی به تعیین موقعیت عکس‌های شبکه اجتماعی فلیکر[۸] پرداخته‌اند. در این پژوهش علاوه بر تحلیل داده-های مکانی از تحلیل محتوای تصاویر نیز بهره برده شده است. در واقع، در این روش ابتدا توزیع مکانی تصاویر اخذشده در محل‌های شاخص[۹] به‌دست می‌آید. سپس با استفاده از روش‌های خوشه‌بندی ارتباط بین محتوای بصری، متنی و ویژگی‌های زمانی با هر یک از خوشه‌ها مشخص می‌شود. با این کار توانایی پیش‌بینی مکان منابع

۱ Ontology
۲ Disambiguation
۳ Toponomy Resolution
۴ Blogs
۵ Geoparser
۶ Tokenization
۷ Part of Speech
۸ Flickr
۹ Landmarks

بهبود می‌یابد. در این تحقیق از ۳۵ میلیون عکس از مکان‌های مهم سرتاسر جهان استفاده شده است. یکی از محدودیت‌های این روش به کار بردن روش‌های خوشه-بندی جهت تعیین موقعیت است. از این رو، دقت تعیین موقعیت به دقت خوشه بندی و تعداد کلاس‌های آن وابسته است. همچنین، این روش برای استفاده در مسائلی که در آن‌ها محل‌های مهم نقشی ندارند، کاربرد ندارد.

در پژوهشی دیگر، فردلند و همکاران [۱۳] به تعیین موقعیت ویدئوهای شبکه اجتماعی فلیکر پرداخته‌اند. این پژوهشگران علاوه بر برچسب‌های هر ویدئو، از محتوای ویدئو نیز برای تعیین موقعیت بهره برده‌اند. روش این پژوهش به این شرح است که ابتدا ویدئوها را به دو دسته آموزشی و تست دسته‌بندی کرده، سپس برچسب‌های ویدئوهای دسته تست را با برچسب‌های دسته آموزشی مقایسه می‌کنند. پس از آن، مرکز هندسی سه مکان ویدئو (آموزشی) با بیشترین شباهت در تعداد برچسب‌ها و کمترین وسعت به عنوان مکان محتمل در نظر گرفته می‌شود. در مرحله بعد، فریم[۱] میانه هر ویدئوی انتخاب شده در مرحله قبل به عنوان ورودی انتخاب می‌شود و عوارض تصویر با استفاده از توصیفگر GIST [۱۴] و بافت‌نگار[۲] رنگی آن استخراج می‌گردد. سپس با محاسبه فاصله بین توصیفگرها و فاصله بین بافت‌نگارها، میزان مشابهت بین تصویر ویدئوی تست و آموزشی به‌دست می‌آید. در نهایت، مختصات مرکز هندسی مکان ویدئوی با کمترین میزان فاصله به عنوان گزینه نهایی انتخاب می‌شود. این روش می‌تواند ۱۴ درصد ویدئوها را با دقت ۱۰ متر تعیین موقعیت کند.

افزون بر زمین‌مرجع‌سازی منابع بصری [۶]، پژوهش-هایی دیگر به زمین‌مرجع‌سازی منابعی همچون صفحات ویکیپدیا[۳] و فرسته[۴]های توئیتر[۵] پرداخته‌اند. برای نمونه، چنگ و همکاران [۱۵] از یک چارچوب احتمالاتی بر اساس براورد درست‌نمایی بیشینه[۶] جهت براورد محل کاربر، در مقیاس شهر، هنگام ارسال یک فرسته در توئیتر بهره برده‌اند. این چارچوب تنها با استفاده از متن فرسته و بدون نیاز به اطلاعات IP[۷] کاربر، اطلاعات ورود یا پایگاه-های دانش خارجی به این امر می‌پردازد. پژوهشگران این تحقیق توانسته‌اند محل ۵۱ درصد کاربران را با دقت ۱۶۰ کیلومتر براورد کنند.

در پژوهشی دیگر، وینگ و بالدریج [۱۶] با استفاده از یک شبکه[۸] بر روی سطح زمین و ایجاد یک مدل زبانی مولد[۹] [۱۷] برای هر بخش این شبکه به براورد موقعیت صفحات ویکیپدیا و فرسته‌های توئیتر پرداخته‌اند. بهترین دقت بدست‌آمده برای صفحات ویکیپدیا با این روش ۱۱/۸ کیلومتر و برای فرسته‌های توئیتر ۴۷۹ کیلومتر است.

۳- مبانی نظری

در این بخش به بررسی مفاهیم، روش‌ها و الگوریتم-های مورد استفاده در این تحقیق می‌پردازیم. ابتدا مفهوم مکان‌مرجع‌سازی به اختصار توضیح داده می‌شود. پس از آن، اطلاعات و داده‌های محل‌مبنا تبیین شده و در انتها الگوریتم جنگل تصادفی شرح داده خواهد شد.

۳-۱- مکان‌مرجع‌سازی

اطلاعات دارای دو ویژگی جدانشدنی «زمان» و «مکان» هستند. به فرایند برقراری ارتباط بین اطلاعات و مکان جغرافیایی آن‌ها مکان‌مرجع‌سازی گفته می‌شود. این ارتباط می‌تواند به صوت غیررسمی، یعنی بصورت اشاره به مکان با استفاده از ساختارهای زبانی، یا بصورت رسمی، یعنی با استفاده از بیان مختصات در یک سیستم مرجع، باشد [۱۸]. پس از فرایند مکان‌مرجع‌سازی، اطلاعاتی که در مورد یک محل بر روی زمین وجود داشته‌اند اما قبلا فاقد مختصات بوده‌اند، دارای مختصات می‌شوند. این اطلاعات شامل طیف گسترده‌ای از انواع داده نظیر تصاویر ماهواره‌ای [۱۹] یا هوایی [۲۰]، آدرس‌ها و کدهای پستی [۲۱]، نقشه‌های تاریخی [۲۲]، منابع متنی داخل کتب و مقالات [۲۳]، منابع رقومی بر بستر شبکه جهانی وب[۱۰] [۲۴] یا حتی گونه‌های جانوری و گیاهی بر روی زمین [۲۵, ۲۶] می‌شود. فرایند مکان‌مرجع‌سازی یک عملیات پایه و اولیه در سیستم‌های اطلاعات مکانی به‌شمار می‌رود و پس از آن می‌توان به انجام محاسبات بر روی اطلاعات پرداخت.

۱ Frame
۲ Histogram
۳ Wikipedia
۴ Post
۵ Twitter
۶ Maximum Likelihood Estimation (MLE)
۷ Internet Protocol (IP)
۸ Grid
۹ Generative Language Model
۱۰ World Wide Web (WWW)

۳-۲- اطلاعات محل‌مبنا

فضا[۱] و محل[۲] دو عبارت بنیادین در حوزه‌های جغرافیا [۲۷, ۲۸]، علوم انسانی [۲۹]، علوم اجتماعی [۳۰] و علوم اطلاعات مکانی[۳] [۳۱] به شمار می‌روند. در این حوزه‌ها، به طور کلی، دیدگاه فضایی به مطالعه هستنده‌ها، عوارض، ویژگی‌ها و رخدادهای سطح زمین یا نزدیک به سطح زمین می‌پردازد. از منظر چنین دیدگاهی، فضا یک محیط پیوسته است. از این رو می‌توان سیستم‌های مختصات مختلفی بر روی آن تعریف کرد. نمونه بارز چنین سیستم مختصاتی، استفاده از طول و عرض جغرافیایی جهت تعیین محل عوارض بر روی سطح زمین است.

از سوی دیگر، محل یک مفهوم برخاسته از جامعه است که در مراودات افراد با یکدیگر شکل می‌گیرد. میان پژوهشگران تعریف محل هنوز مورد مناقشه است [۳۲, ۳۳]. در سال‌های اخیر تمرکز پژوهشگران بر تعریفی از محل است که بیان می‌دارد که یک محل متشکل از ترکیب دو مفهوم فضا و معنا است [۳۴]. با توجه به این که افراد در ارتباط با یکدیگر و برای اشاره به مکان‌ها از مختصات استفاده نمی‌کنند، منظور از فضا در این تعریف یک ارجاع[۴] به موقعیت محل است. به عنوان نمونه، این ارجاع ممکن است به صورت نام منطقه‌ای که محل مورد نظر در آن واقع شده است یا نام یک محل شاخص در نزدیکی محل مورد نظر باشد [۳۵, ۳۶]. به اطلاعاتی که به موقعیتی بر روی زمین اشاره می‌کنند، اما به طور صریح مختصات آن را در بر ندارند، اطلاعات محل‌مبنا گفته می‌شود.

در سال‌های اخیر بخش عظیمی از اطلاعات توسط کاربران شبکه جهانی وب تولید شده اند [۳۷]. در واقع، این روزها کاربران تنها استفاده‌کننده اطلاعات مکانی نیستند، بلکه تولیدکننده آن نیز هستند. علاوه بر این، امروزه کاربران غیرحرفه‌ای و نا آشنا به مفاهیم تخصصی اطلاعات مکانمند نظیر مختصات مرجع از سامانه‌های مکان‌مبنا بهره می‌برند [۳۸]. از این رو، میزان قابل توجهی از داده‌ها حاوی اطلاعات محل‌مبنا هستند. در این میان پژوهش‌هایی وجود دارند که در آن‌ها سعی شده است تا ارتباط میان اطلاعات محل‌مبنا و فضا را ایجاد کنند [۳۹].

یکی از روش‌های برقراری ارتباط میان محل و فضا، مکان-مرجع‌سازی اطلاعات محل‌مبنا است.

۳-۳- الگوریتم جنگل‌های تصادفی

به منظور استفاده از حجم گسترده مجموعه داده‌های در حال تولید در دهه اخیر، به الگوریتم‌هایی نیاز است که قادر به کار با این حجم از اطلاعات هستند و بطور همزمان کارامدی خود را در چنین شرایطی حفظ می‌کنند. الگوریتم‌های یادگیری ماشین[۵] دارای چنین خاصیتی هستند. یکی از شناخته‌شده‌ترین روش‌های یادگیری ماشین الگوریتم جنگل‌های تصادفی [۴۰] است. اساس این الگوریتم بر تفکر تقسیم و حل[۶] قرار دارد، بدین معنا که داده‌ها به بخش‌های مختلفی تقسیم می‌شوند و برای هر بخش یک درخت تصمیم تصادفی ایجاد می‌شود. سپس نتایج تمام درخت‌های ایجاد شده تجمیع شده و پیش‌بینی انجام می‌شود. یکی از دلایل شهرت این الگوریتم قابلیت استفاده از آن در مسائل مختلف است [۴۱-۴۳]. همچنین تعداد پارامترهای موردنیاز برای تنظیم[۷] آن کم است. در ادامه به بررسی نحوه عملکرد این روش خواهیم پرداخت.

فرض کنیم یک مجموعه داده آموزشی D به صورت زیر داشته باشیم:

$$D = \{(\boldsymbol{X}_1, y_1), \dots, (\boldsymbol{X}_n, y_n)\} \quad (۱)$$

که در آن X یک بردار از ویژگی‌ها[۸] و y مقدار مشاهده شده متناظر با بردار X است. همچنین n تعداد نمونه‌هایی است که بصورت تصادفی از اطلاعات برداشت می‌شود. هدف الگوریتم جنگل‌های تصادفی ایجاد یک براوردگر h است که مقدار y را با توجه به مقادیر موجود در X با استفاده از D پیش‌بینی می‌کند.

$$h = \{h_1(\boldsymbol{X}), \dots, h_k(\boldsymbol{X})\} \quad (۲)$$

که در آن k تعداد براوردگرها است. در واقع، در صورتی‌که هر $h_k(X)$ یک درخت تصمیم باشد، الگوریتم به یک مسئله جنگل‌های تصادفی تبدیل می‌شود. این

۱ Space
۲ Place
۳ Geospatial Information Science (GIScience)
۴ Reference
۵ Machine Learning Algorithms
۶ Divide and Conquer
۷ Tuning Parameters
۸ Features

الگوریتم براوردگر نهایی را از تجمیع براوردگرهای رابطه ۲ بدست می‌آورد. اگر فضای پاسخ مسئله گسسته باشد (مسئله طبقه‌بندی)، جواب نهایی آن است که بیشترین تعداد براوردگر به آن رسیده باشند. در صورتی که فضای پاسخ پیوسته باشد (مسئله برازش)، جواب نهایی از طریق میانگین‌گیری جواب‌ها بدست می‌آید.

$$\hat{h} = \frac{1}{k}\sum_{i=1}^{k} h_i(\boldsymbol{X}') \qquad (۳)$$

که در آن $\hat{h}$ براوردگر نهایی و X' بردار نمونه مورد پیش‌بینی است. یکی از ویژگی‌های درخت‌های تصمیم زیاد بودن میزان وردایی[۱] جواب آن‌ها است. جنگل‌های تصادفی روشی برای میانگین‌گیری از درخت‌ها است و از این رو میزان وردایی جواب کاهش می‌یابد [۴۴].

تعیین براوردگر نهایی و بردار نمونه مورد پیش‌بینی بر اساس مجموعه آموزشی مهم‌ترین هدف استفاده از الگوریتم‌های پیش‌بینی است. با این وجود، در برخی موارد، هدف تنها دقیق‌ترین پیش‌بینی نیست، بلکه یافتن ویژگی-هایی است که در فرایند پیش‌بینی بیشترین تاثیر را دارند [۴۵]. یکی از محصولات الگوریتم جنگل‌های تصادفی میزان اهمیت ویژگی‌هایی است که در پیش‌بینی نقش دارند.

۴- پیاده‌سازی و نتایج

این بخش به جزئیات نحوه پیاده‌سازی روش تحقیق اختصاص دارد. در این بخش، ابتدا داده‌های مورد استفاده در تحقیق بررسی و ویژگی‌های آن شرح داده می‌شود. سپس جزئیات پیاده‌سازی مطرح خواهد شد و در نهایت نتایج حاصل از تحقیق ارائه می‌شود.

۴-۱- داده‌های تحقیق

در پژوهش حاضر از آگهی‌های سایت دیوار[۲] در پیاده-سازی روش پیشنهادی استفاده شده است. دیوار یک سکوی خرید و فروش بدون واسطه‌ی برخط، مشابه نمونه خارجی کریگزلیست[۳]، است که از سال ۱۳۹۲ کار خود را در ایران آغاز کرده است. در این سکو کاربران می‌توانند آگهی‌های موردنظر خود را منتشر کنند و محصولات خود را برای فروش ارائه کنند. یکی از آگهی‌های موجود در این پایگاه، املاک و مستغلات است. در این تحقیق، آگهی‌های تبلیغاتی مرتبط با املاک در این وبگاه، شامل اطلاعات موقعیت املاک، نام محله، قیمت، متراژ و اطلاعات توصیفی دیگر، از طریق خزیدن در وب[۴] استخراج شده است. با در نظر گرفتن هدف اصلی این تحقیق، مزیت این داده‌ها نسبت به داده‌های محل‌مبنای دیگر، نظیر آن‌چه که از متون شبکه‌های اجتماعی قابل استخراج است، سهولت و دقت بیشتر در تعیین موقعیت آن‌ها بدلیل عدم نیاز به پردازش زبان طبیعی[۵] و متن‌کاوی[۶] است.

خزیدن در وب مجموعه‌ای از تکنیک‌های مورد استفاده جهت دریافت برخی اطلاعات از یک تارنما[۷] است. در این روش سعی می‌شود تا گردش انسان در فضای اینترنت شبیه‌سازی شود تا بتوان به اطلاعات موجود در یک تارنما دست یافت [۴۶]. این کار بدون اتصال مستقیم به پایگاه داده تارنمای موردنظر و یا اتصال به یک وب سرویس[۸] امکان‌پذیر است. در واقع، صفحات وبی که توسط خادم پردازش شده و از طریق پروتکل HTTP[۹] به سمت مخدوم ارسال شده است، مورد تحلیل قرار می‌گیرند و اطلاعات از آن‌ها استخراج می‌شوند. برای استخراج اطلاعات از روش‌های برنامه‌نویسی بر بستر HTTP، تجزیه سند با استفاده از مفهوم مدل شیءگرای سند (DOM[۱۰]) و تجزیه زبان نشانه‌گذاری ابرمتن (HTML[۱۱]) است.

مزیت استفاده از خزیدن در وب به جای استفاده از وب سرویس‌ها در این است که وب سرویس‌های موجود ممکن است برای تمام داده‌های مورد نیاز موجود نباشند [۴۷]. برای مثال، در تارنمای دیوار هیچ وب سرویسی برای اتصال و دریافت داده وجود ندارد. به‌منظور گرداوری داده‌های آگهی‌های املاک از تارنمای دیوار یک خزنده در وب با استفاده از کتابخانه Scrapy در زبان پایتون ایجاد شد. این خزنده به مدت ۱۴ روز به استخراج آگهی‌های مرتبط با املاک و مستغلات پرداخت و طی این مدت

۱ Variance
۲ https://divar.ir/
۳ https://craigslist.org/

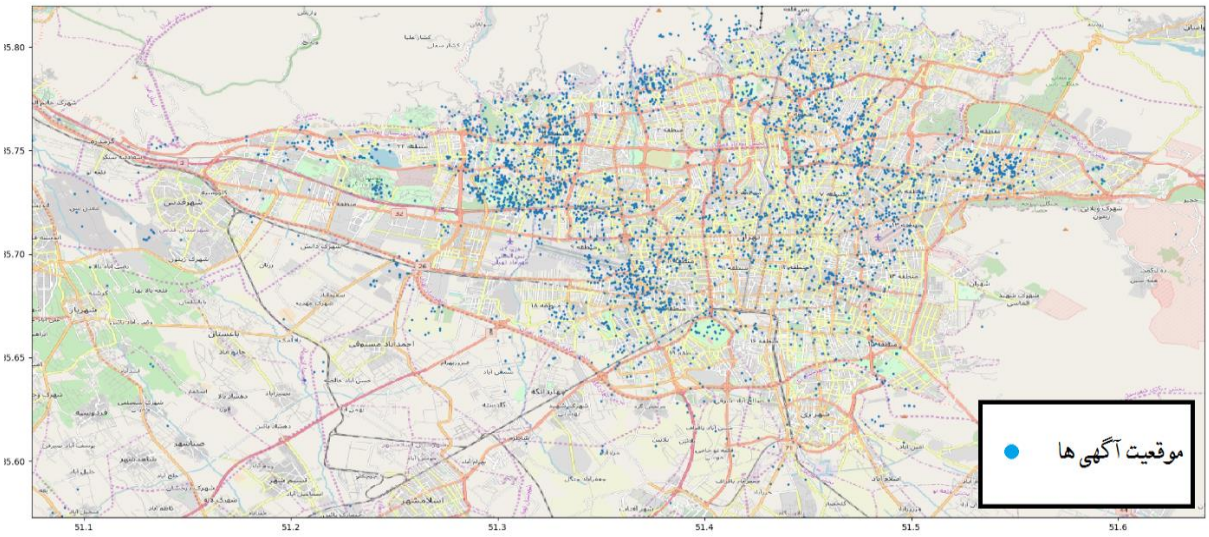
٤ Web Scraping
٥ Natural Language Processing (NLP)
٦ Text Mining
٧ Website
٨ Web Service
٩ Hyper-Text Transfer Protocol (HTTP)
١٠ Document Object Model (DOM)
١١ Hyper-Text Markup Language (HTML)

تعداد ۵۴۷۲۷ آگهی استخراج شد. اطلاعات استخراج شده شامل توضیحات کاربر، نوع آگهی‌دهنده، قیمت اجاره بها، تعداد اتاق‌های ملک، شهری بودن ملک یا واقع شدن در حاشیه شهر، نوع آگهی، سال ساخت، نوع سند، میزان ودیعه، قیمت فروش، مساحت، منطقه‌ای که ملک در ان واقع شده است، طول و عرض جغرافیایی و عنوان است. ذکر این نکته ضروری است که تمام اطلاعات بالا برای تمام آگهی‌ها وجود ندارد. جدول ۱ اطلاعات استخراج شده به همراه مقادیر ممکن برای هر یک را نشان می‌دهد.

جدول ۱- ویژگی‌های استخراج شده از آگهی‌های تبلیغاتی املاک در تارنمای دیوار

ویژگی	توضیحات	تعداد آگهی‌های دارای این ویژگی
توضیحات آگهی	متنی که کاربر جهت تبلیغ ملک خود وارد می‌کند.	۵۴۷۲۷
نوع آگهی‌دهنده	کاربر مشخص می‌کند که آیا فروشنده شخصی است یا مشاور املاک.	۵۴۶۷۰
اجاره‌بها	در صورتی که ملک استیجاری است، مبلغ اجاره بهای ماهانه چه مقدار است؟	۱۶۹۰۴
تعداد اتاق	در صورتی که ملک مسکونی است، تعداد اتاق‌های آن چقدر است؟	۵۰۵۹۴
شهری یا حومه شهری	در صورتی که ملک از نوع زمین است، موقعیت آن در محدوده شهری است یا در حاشیه شهر؟	۲۴۲
نوع آگهی	کاربر مشخص می‌کند که ملک فروشی (زمین، باغ، کارخانه و غیره) است یا استیجاری.	۵۴۷۲۷
سال ساخت	در صورتی که ملک دارای بنا است، بنا در چه سالی ساخته شده است؟	۴۵۱۳۷
نوع سند	کاربر نوع سند (شش دانگ، قولنامه و غیره) را مشخص می‌کند.	۲۰۹۰
میزان ودیعه	در صورتی که ملک استیجاری است، میزان ودیعه را مشخص می‌کند.	۱۶۹۰۷
قیمت فروش	در صورتی که ملک فروشی است، مبلغ فروش چقدر است؟	۳۶۲۰۴
تبادل با ملک دیگر	اگر کاربر مایل به تبادل با ملک دیگر باشد، در این بخش تعیین می‌کند.	۳۳۱
مساحت	مساحت ملک چقدر است؟	۵۳۴۴۲
منطقه	ملک در کدام منطقه از تهران واقع شده است؟	۵۴۷۲۷
طول جغرافیایی	طول جغرافیایی موقعیت ملک را مشخص می‌کند.	۵۹۰۹
عرض جغرافیایی	عرض جغرافیایی موقعیت ملک را مشخص می‌کند.	۵۹۰۹
عنوان	عنوان آگهی تبلیغاتی را مشخص می‌کند.	۵۴۷۲۷

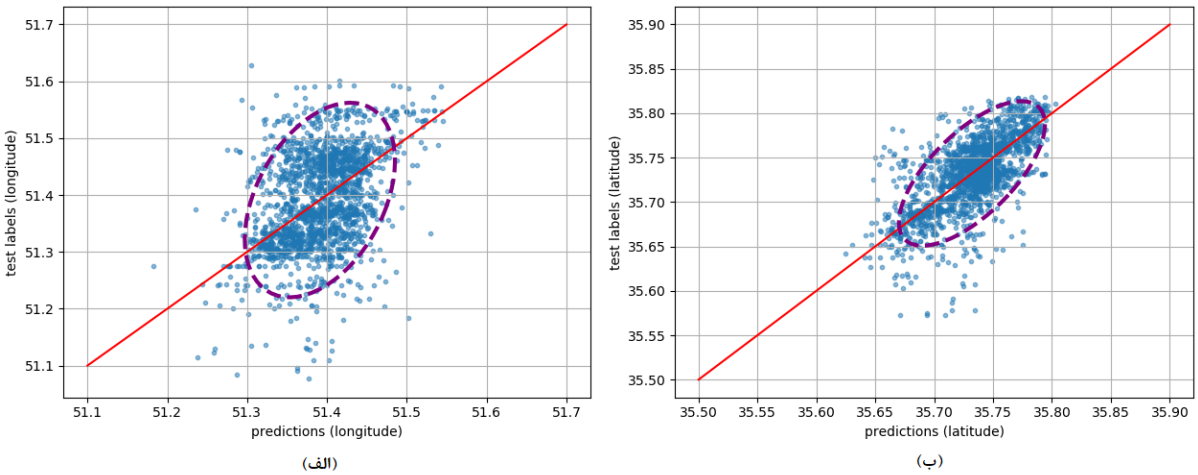
همانطور که در جدول ۱ مشخص است، از میان آگهی‌های استخراج شده تعداد ۵۹۰۹ آگهی دارای مختصات‌اند و برای آموزش الگوریتم موردنظر مناسب هستند. همچنین، برخی از موقعیت‌ها اشتباه هستند و یا خارج از شهر تهران قرار دارند. از این رو، پس از پیش-پردازش داده‌ها، تعداد ۵۴۴۵ آگهی برای ادامه کار انتخاب شدند. علاوه بر این، برخی ویژگی‌ها کارایی لازم برای استفاده در این الگوریتم را ندارند. برای مثال، ویژگی‌های نوع سند، توضیحات، عنوان، نوع آگهی‌دهنده و نوع آگهی الگوی مشخص برای آموزش الگوریتم را ندارند؛ از این رو، این ویژگی‌ها از داده‌های آموزشی کنار گذاشته شدند. با توجه به این‌که دقت این روش کمتر از وسعت محله‌های شهر تهران است، اطلاعات محله‌ها نیز از مجموعه داده‌ها حذف شدند. در واقع، در این پژوهش موقعیت آگهی‌های تبلیغاتی املاک با استفاده از قیمت اجاره بهاء و قیمت فروش و همچنین سال ساخت آن پیش‌بینی می‌شود. پراکندگی داده‌های تحقیق در شکل ۱ نشان داده شده است.



شکل ۱- پراکندگی آگهی‌های تبلیغاتی استخراج‌شده (داده‌های تحقیق) در سطح شهر تهران و حومه

۴-۲- نتایج و بحث

به منظور پیاده‌سازی الگوریتم جنگل‌های تصادفی داده‌ها به دو دسته آموزشی و تست تقسیم شدند. داده‌های آموزشی ۶۰ درصد داده‌های کل و داده‌های تست ۴۰ درصد داده‌های کل را در بر می‌گیرند. پس از انجام عملیات سعی و خطا، با توجه به حجم داده‌های آموزشی، بهترین دقت هنگامی حاصل می‌شود که تعداد درخت‌های الگوریتم جنگل‌های تصادفی ۱۰۰ عدد در نظر گرفته شود. با اجرای الگوریتم، دقت به دست آمده در پیش‌بینی موقعیت‌های آگهی‌های دسته تست برابر با ۶/۶ کیلومتر در راستای طول جغرافیایی و ۲/۱ کیلومتر در راستای عرض جغرافیایی است. این در حالی است که وسعت مناطق ۲۲ گانه شهر تهران به طور میانگین برابر ۱۶ کیلومتر مربع است. بنابراین، قابلیت پیش‌بینی منطقه ملک وجود دارد. در شکل ۲ (الف) و (ب) بیضی خطای حاصل از نتایج به ترتیب بر روی نمودار نقطه‌ای پیش‌بینی‌های طول و عرض جغرافیایی نشان داده شده است. در این شکل‌ها محور افقی نشان‌دهنده مختصات پیش‌بینی‌شده و محور قائم نشان دهنده مختصات داده‌های تست است. همچنین، خط همانی با رنگ قرمز نشان داده شده است.



شکل ۲- (الف) نمودار نقطه‌ای پیش‌بینی‌های طول جغرافیایی؛ (ب) نمودار نقطه‌ای پیش‌بینی‌های عرض جغرافیایی

بیضی‌های خطا در سطح اطمینان ۹۰ درصد رسم شده‌اند. در این نمودارها، هر چه ابر نقاط به خط همانی نزدیک‌تر باشند، دقت پیش‌بینی‌ها بیشتر است. به عبارت دیگر، هر چه کشیدگی بیضی خطا در جهت خط همانی بیشتر باشد، دقت بیشتر است. همانطور که از شکل‌ها مشخص است، دقت پیش‌بینی مولفه عرض جغرافیایی

نسبت به مولفه طول جغرافیایی بیشتر است. همانطور که در بخش مبانی نظری بیان شد، با استفاده از الگوریتم جنگل‌های تصادفی می‌توان به تعیین میزان اهمیت ویژگی‌ها در پیش‌بینی نیز پرداخت. جدول‌های ۲ و ۳ میزان اهمیت متغیرهای پیش‌بینی در پیش‌بینی محل آگهی را نشان می‌دهد.

همانطور که از جدول‌های ۲ و ۳ مشخص است، قیمت ملک و میزان ودیعه و اجاره‌بها بیشترین تاثیر را در تعیین موقعیت دارد. پس از آن مساحت ملک و پس از آن سال ساخت آن در تعیین موقعیت موثر هستند. همچنین، از این جداول می‌توان دریافت که قیمت مسکن در راستای شمالی – جنوبی شهر تهران دارای الگوی بیشتری نسبت به راستای شرقی – غربی است.

جدول ۲- درصد اهمیت نسبی متغیرهای پیش‌بینی (طول جغرافیایی)

ویژگی	اهمیت نسبی
قیمت / ودیعه + اجاره بهاء	٪۴۲
مساحت	٪۳۷
سال ساخت	٪۲۱

جدول ۳- درصد اهمیت نسبی متغیرهای پیش‌بینی (عرض جغرافیایی)

ویژگی	اهمیت نسبی
قیمت / ودیعه + اجاره بهاء	٪۵۳
مساحت	٪۳۱
سال ساخت	٪۱۶

۵- نتیجه‌گیری

تعیین موقعیت منابع مختلف یکی از فعالیت‌های پرکاربرد در علوم اطلاعات مکانی به‌شمار می‌رود. با گسترش فعالیت افراد در فضای مجازی میزان انتشار منابع تحت وب در سال‌های اخیر افزایش یافته است. یکی از این منابع، آگهی‌های تبلیغاتی املاک است. در این تحقیق، با استفاده از الگوریتم یادگیری ماشین جنگل‌های تصادفی به مکان‌مرجع‌سازی آگهی‌های تبلیغاتی پرداخته شده است. نتایج این تحقیق نشان می‌دهد که این روش قابلیت تعیین موقعیت آگهی‌های تبلیغاتی در سطح مناطق تهران را دارا می‌باشد. علاوه بر این، نتایج این تحقیق به خوبی نشان می‌دهد که قیمت املاک در سطح شهر تهران در راستای شمالی جنوبی نسبت به راستای شرقی غربی دارای الگوی بیشتری است. همچنین، می‌توان دریافت که قیمت مسکن بیشترین تاثیر را در تعیین موقعیت آگهی‌ها ایفا می‌کند.

یکی از محدودیت‌های این روش عدم پویا بودن آن است. در واقع، زمانی‌که قیمت املاک در طول زمان تغییر کند، این الگوریتم قادر به سازگاری با قیمت‌های جدید نیست، مگر اینکه داده‌های جدید آموزشی به آن تزریق شود. این موضوع بخصوص برای کشورهایی نظیر ایران که میزان تغییرات قیمت مسکن در آن‌ها در طول زمان بسیار است، چالشی بزرگ محسوب می‌شود. یکی از پیشنهادها برای تحقیقات آتی برطرف کردن این معضل است. علاوه بر این، تحقیقات در آینده باید به این موضوع بپردازند که دقت این روش برای شهرهایی که تفاوت قیمت املاک در آن‌ها زیاد است، به چه صورت است.

مراجع

[1] O. Serrat, "Social network analysis," in Knowledge solutions, ed: Springer, 2017, pp. 39-43.

[2] T. Blaschke, H. Merschdorf, P. Cabrera-Barona, S. Gao, E. Papadakis, and A. Kovacs-Györi, "Place versus Space: From Points, Lines and Polygons in GIS to Place-Based Representations Reflecting Language and Culture," ISPRS International Journal of Geo-Information, vol. 7, p. 452, 2018.

[3] J. Pereira, J. Monteiro, J. Estima, and B. Martins, "Assessing flood severity from georeferenced photos," in Proceedings of the 13th Workshop on Geographic Information Retrieval, 2019, pp. 1-10.

[4] R. S. Purves, P. Clough, C. B. Jones, M. H. Hall, and V. Murdock, "Geographic information retrieval: progress and challenges in spatial search of text," Foundations and Trends ®in Information Retrieval, vol. 12, pp. 164-318, 2018.

[5] S. Serra, A. Jorge, and T. Chambel, "Multimodal Access to Georeferenced Mobile Video through Shape, Speed and Time," in Proceedings of the 28th International BCS Human Computer Interaction Conference (HCI 2014) 28, 2014, pp. 347-352.

[6] O. Van Laere, S. Schockaert, and B. Dhoedt, "Georeferencing Flickr resources based on textual meta-data," Information Sciences, vol. 238, pp. 52-74, 2013.

[7] D. Dias, I. Anastácio, and B. Martins, "A language modeling approach for georeferencing textual documents," in Actas del Congreso Español de Recuperación de Información, 2012.

[8] D. Buscaldi and P. Rosso, "Geo-WordNet: Automatic Georeferencing of WordNet," in LREC, 2008.

[9] C. Fink, C. Piatko, J. Mayfield ,D. Chou, T. Finin, and J. Martineau, "The geolocation of web logs from textual clues," in 2009 International Conference on Computational Science and Engineering, 2009, pp. 1088-1092.

[10] H. Chen, S. Winter, and M. Vasardani, "Georeferencing places from collective human descriptions using place graphs," Journal of Spatial Information Science, vol. 2018, pp. 31-62, 2018.

[11] C. Grover, R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, et al., "Use of the Edinburgh geoparser for georeferencing digitized historical collections," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 368, pp. 3875-3889, 2010.

[12] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos ",in Proceedings of the 18th international conference on World wide web, 2009, pp. 761-770.

[13] G. Friedland, J. Choi, and A. Janin, "Video2GPS: a demo of multimodal location estimation on flickr videos," in ACM Multimedia, 2011, pp. 833-834.

[14] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International journal of computer vision, vol. 42, pp. 145-175, 2001.

[15] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 759-768.

[16] B. P. Wing and J. Baldridge, "Simple supervised document geolocation with geodesic grids ",in Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1, 2011, pp. 955-964.

[17] C. De Rouck, O. Van Laere, S. Schockaert, and B. Dhoedt, "Georeferencing Wikipedia pages using language models from Flickr," in 10th International Semantic Web Conference (ISWC 2011), 2011.

[18] L. L. Hill, Georeferencing: The geographic associations of information: Mit Press, 2009.

[19] T. Long, W. Jiao, G. He, and Z. Zhang, "A fast and reliable matching method for automated georeferencing of remotely-sensed imagery," Remote sensing, vol. 8, p. 56, 2016.

[20] G. Forlani, F. Diotri, U. M. d. Cella, and R. Roncella, "Indirect UAV strip georeferencing by on-board GNSS data under poor satellite coverage," Remote Sensing, vol. 11, p. 1765, 2019.

[21] P. Norman, "Georeferencing socio-demographic information: postcodes, addresses, land parcels and geographical information systems," 2016.

[22] M. Gullu and O. G. Narin, "Georeferencing of the Nile River in Piri Reis 1521 map, Using Artificial Neural Network Method," Acta Geodaetica et Geophysica, pp. 1-15, 2019.

[23] K. McDonough, L. Moncla, and M. Van de Camp, "Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora," International Journal of Geographical Information Science, pp. 1-25, 2019.

[24] A. Tabarcea, V. Hautamäki, and P. Fränti, "Ad-hoc georeferencing of web-pages using street-name prefix trees," in International Conference on Web Information Systems and Technologies, 2010, pp. 259-271.

[25] T. D. Bloom, A. Flower, and E. G. DeChaine, "Why georeferencing matters: Introducing a practical protocol to prepare species occurrence records for spatial analysis," Ecology and evolution, vol. 8, pp. 765-777, 2018

[26] C. L. Kelloff and L. B. Kass, "DATABASING AND GEOREFERENCING HISTORICAL COLLECTIONS TO DISCOVER POTENTIAL SITES FOR RARE AND ENDANGERED PLANTS OF NEW YORK, USA," Journal of the Botanical Research Institute of Texas, vol. 12, 2018.

[27] H. Couclelis, "Space, time, geography," Geographical information systems, vol. 1, pp. 29-38, 1999.

[28] T. Cresswell, "Place: encountering geography as philosophy," Geography, vol. 93, pp. 132-139, 2008.

[29] D. J. Bodenhamer, "The spatial humanities: space, time and place in the new digital age," in History in the Digital Age, ed: Routledge, 2012, pp. 35-50.

[30] S. J. Steinberg and S. L. Steinberg, Geographic information systems for the social sciences: investigating space and place: Sage Publications, 2005.

[31] S. Winter, W. Kuhn, and A. Krüger, "Guest editorial: does place have a place in geographic information science?," ed: Taylor & Francis, 2009.

[32] D. Massey, A global sense of place: Aughty. org, 2010.

[33] P. Agarwal, "Contested nature of place: Knowledge mapping for resolving ontological distinctions between geographical concepts," in International Conference on Geographic Information Science, 2004, pp. 1-21.

[34] J. A. Agnew, "A theory of place and politics," Place and politics, 1987.

[35] G. McKenzie, Z. Liu, Y. Hu, and M. Lee, "Identifying urban neighborhood names through user-contributed online property listings," ISPRS International Journal of Geo-Information, vol. 7, p. 388, 2018.

[36] H. Chen, M. Vasardani, and S. Winter, "Clustering-based disambiguation of fine-grained place names from descriptions," GeoInformatica, pp. 1-24, 2019.

[37] M. F. Goodchild, "Citizens as sensors: the world of volunteered geography," GeoJournal, vol. 69, pp. 211-221, 2007.

[38] A. Turner, Introduction to neogeography: " O'Reilly Media, Inc.", 2006.

[39] E. Papadakis, G. Baryannis, and T. Blaschke, "From space to place and back again: towards an interface between space and place," in Platial'18: Workshop on Platial Analysis, 2018.

[40] L. Breiman, "Random forests," Machine learning, vol. 45, pp. 5-32, 2001.

[41] H. Sun, D. Gui, B. Yan, Y. Liu, W. Liao, Y. Zhu, et al., "Assessing the potential of random forest method for estimating solar radiation using air pollution index," Energy Conversion and Management, vol. 119 ,pp. 121-129, 2016.

[42] Y. Qi, "Random forest for bioinformatics," in Ensemble machine learning, ed: Springer, 2012, pp. 307-323.

[43] Z. Lei, T. Fang, H. Huo, and D. Li, "Rotation-invariant object detection of remotely sensed images based on texton forest and hough voting," IEEE Transactions on Geoscience and Remote Sensing, vol. 50, pp. 1206-1217, 2011.

[44] J. Friedman, T. Hastie, and R. Tibshirani, The elements of statistical learning vol. 1: Springer series in statistics New York, 2001.

[45] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in Advances in neural information processing systems, 2013, pp. 431-439.

[46] E. Vargiu and M. Urru, "Exploiting web scraping in a collaborative filtering-based approach to web advertising," Artif. Intell. Research, vol. 2, pp. 44-54, 2013.

[47] D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato, and F. Fdez-Riverola, "Web scraping technologies in an API world," Briefings in bioinformatics, vol. 15, pp. 788-797, 2013.