نشریه علمی علوم و فنون نقشه برداری، دوره یازدهم، شماره ۲، آذر ماه ۱۴۰۰ – مقاله پژوهشی

غنی‌سازی جاینامه با استفاده از آگهی‌های املاک

مهدی شاخصی[*1]، علی‌اصغر آل‌شیخ[2]، رویا حبیبی[3]

[1] دانشجوی کارشناسی ارشد سیستم‌های اطلاعات مکانی - دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه نصیرالدین طوسی
m_shakhesi@email.kntu.ac.ir

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

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

(تاریخ دریافت بهمن ۱۳۹۹، تاریخ تصویب فروردین ۱۴۰۰)

چکیده

با توجه به افزایش روزافزون کاربردهای بازیابی اطلاعات مکانی، جاینامه‌ها به عنوان بخش مهمی از فرآیند بازیابی اطلاعات مکانی، نیازمند غنی‌سازی هستند. یکی از جنبه‌های غنی‌سازی شناسایی و افزودن نام‌های جغرافیایی جدید به جاینامه و به‌هنگام‌سازی آن می‌باشد. از جمله چالش‌های مهم در غنی‌سازی جاینامه‌ها، در نظر گرفته شدن دیدگاه رسمی و اغلب نادیده گرفته شدن جاینام های محلی و همچنین پرهزینه و زمانبر بودن به‌هنگام‌سازی جاینامه ها است. در این تحقیق، با تمرکز بر گردآوری جاینام‌های شهری، روشی داده محور جهت شناسایی نام‌های جغرافیایی از نوع همسایگی و خیابان با استفاده از آگهی‌های املاک ارائه شده است. آگهی‌های املاک برای چهار کلانشهر تهران، مشهد، اصفهان و شیراز از وبسایت دیوار وبکاوی شده و پس از استخراج ان-گرم‌ها و اعمال پیش‌پردازش‌های لازم، ان-گرم‌ها برچسب‌گذاری شدند. بر مبنای ۲۴ معیار مکانی و تحت مدل جنگل تصادفی برای هر کدام از این چهار شهر مدل تولید شده و روی داده سایر شهرها آزموده شد. نتایج نشان‌دهنده‌ی این است که هم در شناسایی خیابان و هم همسایگی، عملکرد مدل آموزش‌یافته براساس داده شهر اول و آزمون روی داده سایر شهرها قابل قبول است. برای مثال، مدل آموزش یافته براساس داده شهر تهران در آزمون روی شهر مشهد، مقادیر ۶۱٪ و ۷۴٪ را برای F_score به ترتیب در شناسایی خیابان و همسایگی کسب کرده است. لذا بر این اساس می‌توان گفت که گردآوری نام‌های جغرافیایی در شرایطی که ابزارهای پردازش متن از کارایی کافی برخوردار نباشند، می‌تواند با تکیه بر رفتار مکانی آن‌ها به خوبی انجام پذیرد.

واژگان کلیدی: غنی‌سازی جاینامه، بازیابی اطلاعات مکانی ، آگهی‌های املاک، جنگل‌های تصادفی

* نویسنده رابط

۱- مقدمه

جاینامه به‌عنوان مجموعه‌ای از نام‌های جغرافیایی شناخته می‌شود که در ساده‌ترین حالت بایستی شامل سه‌تایی «نام، مختصات و دسته‌بندی» به ازای هر مکان بوده باشد [۱]. فرهنگستان زبان و ادب فارسی واژه جاینامه را معادل واژه Gazetteer در نظر گرفته است و آن را «*کتابی مشتمل بر سیاهه اسامی اماکن، همراه با اطلاعات توصیفی و جغرافیایی و تاریخی و آماری*» تعریف کرده است. با این وجود، با توجه به تحولاتی که در حوزه کامپیوتر و فناوری اطلاعات اتفاق افتاده است، جاینامه را بیشتر با عناوینی نظیر پایگاه داده جغرافیایی [۲]، پایگاه دانش جغرافیایی [۳] و سیستم سازمان‌دهنده دانش معرفی می‌کنند [۴].

کاربردهای اصلی جاینامه‌ها در بازیابی اطلاعات جغرافیایی است که عبارت‌اند از: ۱- شناسایی جاینام‌ها در متن [۵-۷]؛ ۲- رفع ابهام جاینام‌های احتمالی[۱] [۸]؛ ۳- نگاشت جاینام به مختصات[۲] [۹]؛ ۴- ارائه لیستی از مکان‌هایی که در محدوده جغرافیایی مشخص‌شده‌ای قرار دارند[۳] [۱۰] و ۵- پاسخ‌دهی به پرسش‌های پیچیده‌تر مکانی. علاوه‌بر این موارد می‌توان به کاربرد جاینامه در نقشه‌خوانی هم اشاره کرد [۱۱].

با توجه به نیازهای جامعه، تعریف پایه جاینامه از سه‌تایی «نام، مختصات و دسته‌بندی» مکان فراتر می‌رود [۱] و به دنبال آن موضوع غنی‌سازی جاینامه مطرح می‌شود. یکی از جنبه‌های غنی‌سازی، افزودن نام‌های جغرافیایی جدید به جاینامه است. با توجه به این که جاینامه‌ها توسط سازمان‌های رسمی تهیه می‌شوند، از دو نظر جاینامه شامل همه نام‌های جغرافیایی مورد استفاده در گستره جغرافیایی تعریف‌شده برای جاینامه نمی‌شود: ۱- دیدگاه رسمی و از بالا به پایین اتخاذ شده در توسعه باعث می‌شود جاینام‌های محلی که توسط خود مردم در گفت‌وگوهای روزمره استفاده می‌شود، در این فرآیند در نظر گرفته نشوند [۱۲, ۱۳] و ۲- زمان‌بر بودن روند به‌هنگام‌سازی جاینامه در کنار سایر هزینه‌ها نیز سبب می‌شود جاینام‌های نوظهور مربوط به مکان‌های تازه تاسیس، حضوری در مجموعه نام‌های جغرافیایی جاینامه نداشته باشند [۱۴].

آنچه که باید یادآور شد تفاوت مکان[۴] و عارضه (جغرافیایی)[۵] است. مکان یعنی موقعیتی جغرافیایی که به عنوان برساختی اجتماعی[۶] شناخته می‌شود. در حالی که عارضه عنصر فیزیکی منحصر به فردی است مثل ساختمان، دریاچه و یا کوه با موقعیت و محدوده جغرافیایی مشخص. تقسیمات سیاسی معمولاً به عنوان عارضه در نظر گرفته می‌شوند چون دارای موقعیت و محدوده جغرافیایی مشخص‌شده‌ای هستند. مفهوم مکان علاوه‌بر عارضه، موقعیت‌هایی را نیز شامل می‌شود که محدوده مبهمی دارند مثل مرکز شهر. آنچه که در تهیه و توسعه جاینامه‌های رسمی از آن صرف نظر می‌شود همین مکان‌های با محدوده و شاید با موقعیت مبهم است [۱۵].

لزوماً هر نقطه‌ای بر روی زمین به عنوان مکان در نظر گرفته نمی‌شود و این اهمیت آن نقطه است که باعث می‌شود دارای نام مختص خود باشد. همچنان که ممکن است مکانی دارای چندین نام جغرافیایی باشد (مثل شهسوار و تنکابن که هر دو به یک مکان اشاره دارند)، یک نام جغرافیایی نیز می‌تواند به بیش از یک مکان اشاره داشته باشد (مثل هادیشهر که نام دو شهر یکی در استان آذربایجان‌شرقی و دیگری در استان مازندران است) [۱۵].

مولفه دوم یعنی مختصات بر مبنای قدرت تفکیک در نظر گرفته‌شده و نیز نوع عارضه جغرافیایی می‌تواند به یکی از پنج حالت نقطه، کادر محصورکننده[۷]، خط، چندضلعی و یا نمایش شبکه‌ای نمایش داده شود [۱]. دسته‌بندی مکان‌ها براساس کاربردی که برای جاینامه لحاظ شده است می‌تواند متفاوت باشد. برای جاینامی مثل «تهران»، دسته‌بندی یا نوع عارضه جغرافیایی همان «شهر» خواهد بود. نوع عارضه جغرافیایی می‌تواند حتی از روی نام جغرافیایی منتسب به آن اخذ شود (فرودگاه امام خمینی).

---

۱ هنگامی که شباهتی بین یک نام جغرافیایی و یک نام غیرجغرافیایی باشد و یا دو مکان نام جغرافیایی مشابهی داشته باشند.

۲ آنچه که در سرویس‌های نقشه مبتنی بر وب یا سیستم‌های ناوبری اتفاق می‌افتد که تحت عنوان Geocoding شناخته می‌شود.

۳ ساده‌ترین مثالی که می‌توان زد شاید جستجوی رستوران‌های یک محله باشد که معکوس کاربرد قبلی است (Reverse Geocoding).

٤ Place

٥ Feature

٦ Social construct؛ ایده‌ای است که توسط جامعه مطرح شده و مورد پذیرش قرار گرفته است.

٧ Bounding Box

نام‌های جغرافیایی جدید را می‌توان از منابع مختلفی گردآوری کرد که از آن جمله می‌توان به شبکه‌های اجتماعی و اطلاعات مکانی مردم گستر اشاره کرد. نام‌های جغرافیایی حتی ممکن است به کمک موتورهای جستوجوگر وب اخذ شوند. توجه ویژه‌ای که به استفاده از این منابع می‌شود، به خاطر دو دلیل اصلی به هنگام بودن و شامل بودن جاینام‌های محلی است [۱۴].

هدف این تحقیق ارائه روشی داده محور برای گردآوری نام‌های جغرافیایی با استفاده از آگهی‌های املاک منتشرشده در وبسایت‌های املاک بوده است. کارکرد روش ارائه‌شده برای جاینام‌هایی است که به خاطر شناخته‌شده بودن در یک منطقه، بصورت مستقل و بدون همراهی واژه‌هایی برای مشخص کردن مکانی بودن آن‌ها در متن آگهی نوشته می‌شوند (نام همسایگی‌ها و خیابان‌های اصلی). دو سوال اصلی پژوهش این بوده است که ۱- با توجه به رفتار مکانی این گونه جاینام‌ها، آیا می‌توان آن‌ها را بر اساس معیارهای مکانی استخراج کرد؟ و ۲- آیا بر مبنای مدل تولیدشده بر اساس داده یک شهر می‌توان همسایگی‌ها و خیابان‌های اصلی سایر شهرها را مشخص نمود؟

استفاده از آگهی‌های املاک در گردآوری جاینام‌ها همان مزایایی را به همراه دارد که برای اطلاعات مکانی مردم گستر و شبکه‌های اجتماعی مطرح می‌شود. با این وجود، دامنه نام‌های جغرافیایی آگهی‌های املاک تا حدی متفاوت‌تر از این گونه منابع است. نمونه بارز این تفاوت در ارجاع‌دهی به نوع مکان است. به بیان دیگر، برخلاف سایر محیط‌های مشارکتی مثل Flickr که بر روی جاینام‌های با محوریت گردشگری تمرکز دارند، در آگهی‌های املاک محوریت با سکونت و ارتباط است که به ترتیب بر روی دو دسته‌بندی همسایگی و خیابان معطوف می‌شود. لذا در این پژوهش، گردآوری جاینام‌های شهری از نوع همسایگی و خیابان از آگهی‌های املاک مدنظر بوده است.

در ادامه، در بخش دوم با مروری بر پیشینه تحقیق روش‌های ارائه‌شده برای گردآوری نام‌های جغرافیایی موردمطالعه قرار گرفته‌اند. در بخش سوم، روش توسعه داده‌شده ارائه گردیده است. در بخش چهارم، نتایج حاصل از پیاده‌سازی این روش ارائه شده و مورد ارزیابی قرار گرفته‌اند. در نهایت، نتیجه‌گیری‌ها و محدودیت‌ها و پیشنهادات برای مطالعات آینده در بخش پنجم آورده شده است.

۲- پیشینه تحقیق

غنی‌سازی جاینامه از جوانب مختلفی مطرح می‌شود: افزودن جاینام‌های جدید [۱۴, ۱۶, ۱۷]، بهبود مولفه مکانی [۱۸, ۱۹]، ایجاد ساختاری معنایی برای جاینامه [۲۰-۲۲]، حذف موارد تکراری [۲۳-۲۵]، افزودن مولفه زمان [۲۶, ۲۷] و حتی چندزبانی کردن جاینامه [۲۸] از جمله این موارد است. با این وجود، در این پژوهش تنها جنبه اول غنی‌سازی جاینامه یعنی افزودن نام‌های جغرافیایی جدید از طریق گردآوری آن‌ها از منابع دیگر، مدنظر بوده است. لذا در ادامه، تنها روش‌های اتخاذشده در پژوهش‌های پیشین مربوط به این جنبه از غنی‌سازی مرور خواهد شد.

گردآوری جاینام‌های جدید و افزودن آن‌ها به جاینامه می‌تواند به روش‌های مختلفی انجام شود. متداول‌ترین روش، استفاده از سایر جاینامه‌ها است. از جاینامه‌های مطرحی که به این شیوه ساخته شده است، جاینامه کتابخانه دیجیتال اسکندریه[۱] است که حاصل تلفیق دو جاینامه GNS[۲] و GNIS[۳] است که به ترتیب توسط آژانس ملی اطلاعات مکانی و سازمان زمین‌شناسی آمریکا توسعه یافته‌اند [۱]. مجموعه نام‌های جغرافیایی اخذ شده به این روش بیشتر به عنوان سنگ‌بنای سایر روش‌ها در نظر گرفته می‌شود. از این رو، در بیشتر پژوهش‌های انجام گرفته که استفاده از سایر روش‌ها مدنظرشان بوده همچنان این روش نیز استفاده شده است. با ظهور وب ۲، استفاده از اطلاعات جعرافیایی مردم گستر، شبکه‌های اجتماعی و در کل، محیط‌هایی که برمبنای مشارکت مردم توسعه یافته‌اند، در چندین پژوهش به عنوان راهکاری برای غنی‌سازی جاینامه مطرح شد. به هنگام‌سازی سریع‌تر جاینامه و نیز اتخاذ سیاستی پایین به بالا در گردآوری نام‌های جغرافیایی دو مزیت مهم این روش محسوب می‌شود. علاوه‌بر دو روش گفته‌شده، بهره‌گیری از موتورهای جستجوگر وب نیز مطرح می‌شود که در برخی مطالعات انجام گرفته بیشتر با دیدگاه حاکم بر روش دوم (استفاده از محتوای کاربرساخته) همراه بوده‌اند [۲۹].

برخلاف جاینامه‌های قبلی که تنها توسط سازمانی رسمی تهیه و نگهداری می‌شدند، در نسل جدید جاینامه‌ها

---

۱ Alexandria Digital Library (ADL)
۲ GEOnet Names Server
۳ Geographic Names Information System

این نگاه صرفاً رسمی بسط یافته به نحوی که استفاده از اطلاعات مردم گستر را نیز شامل می‌شود [۳]. مسلماً جاینامه‌ای که تنها بر مبنای نام‌های جغرافیایی رسمی است در حل بحران‌ها و مسائلی که مشارکت مردم را می‌طلبد کارآمدی آنچنانی نخواهد داشت. لذا یکی از منابع بالقوه گردآوری نام‌های جغرافیایی اطلاعات مردم گستر است. این امر به دو صورت انجام می‌شود. در حالت اول، اطلاعات جغرافیایی از وبسایت‌هایی مثل Wikipedia، Wikimapia، OpenStreetMap اخذ می‌شود و پس از اعمال تغییرات لازم و پیاده‌سازی استانداردهای درنظرگرفته‌شده برای جاینامه، نام‌های جغرافیایی به جاینامه اضافه می‌شوند [۳۰-۳۲]. در حالت دوم، تهیه‌کننده جاینامه با ساخت رابطی بر بستر وب، از ساکنان منطقه مورد نظر درخواست مشارکت می‌کند. اطلاعات مشارکتی نهایتاً بصورت جاینام و سایر اطلاعات جغرافیایی در سامانه ذخیره می‌شود [۳۳].

علاوه‌بر اطلاعات مردم‌گستر، شبکه‌های اجتماعی نیز می‌تواند منبع گردآوری نام‌های جغرافیایی باشد. برای مثال، Lim و همکاران روشی را ارائه کرده‌اند که ایجاد و به‌هنگام‌رسانی جاینامه از طریق توئیت‌های منتشرشده (که دارای برچسب مکانی بوده‌اند) متناسب با بازه‌های زمانی انجام گرفته است به نحوی که سامانه طراحی‌شده قادر باشد محتواهای تولیدشده توسط بات‌ها را شناسایی کند و پالایش نماید [۳۴]. در تحقیقی که توسط Gao و همکاران انجام شده است، بستری توزیع‌یافته و مقیاس‌پذیر ساخته شده و کلان داده شبکه اجتماعی Flickr در یک زیست‌بوم هدوپ[۱] پردازش شده و جاینام‌های جدید شناسایی شده‌اند[۳۵].

در برخی از تحقیقات، گردآوری نام‌های جغرافیایی جدید به کمک موتورهای جستجوگر وب انجام گرفته است. برای نمونه، Zhang و همکاران، چارچوبی مبتنی بر یادگیری ماشین توسعه داده‌اند که به کمک موتور جستجوگر گوگل، جاینام‌های شهری از صفحات وب گردآوری می‌شود [۱۲, ۱۷].

دو پژوهشی که به نوعی مبنای این تحقیق محسوب می‌شوند، شناسایی و گردآوری جاینام‌های جدید از میان آگهی‌های املاک را معرفی کرده‌اند. McKenzie و همکاران با روشی که ارائه کرده‌اند نام همسایگی‌ها برای سه شهر واشنگتن، سیاتل و مونترال از میان آگهی‌ها شناسایی شده است؛ از هر آگهی، ان-گرم‌هایی[۲] استخراج شده است و در نهایت، برای هر ان-گرم، مجموعه مختصات جغرافیایی آگهی‌های شامل آن ان-گرم اختصاص یافته است. برمبنای شاخصه‌های مکانی تعریف‌شده و به کمک یادگیری ماشین، همسایگی‌های جدیدی شناسایی شده است [۳۶]. ولی در روشی که Hu و همکاران ارائه کرده‌اند، به کمک ابزارهای پردازش زبان طبیعی، جاینام‌های احتمالی شناسایی شده و در ادامه، با خوشه‌بندی مکانی جاینام‌های نهایی مشخص شده است [۱۴].

۳- روش توسعه داده شده

در این بخش، ابتدا روش توسعه داده شده (شکل ۱) به اختصار معرفی شده و در ادامه، مراحل کار به تفصیل بررسی شده و چگونگی پیاده‌سازی آن‌ها نیز ارائه شده است.

علیرغم پیشرفت‌های صورت‌گرفته در سال‌های اخیر در توسعه ابزارهای پردازش زبان فارسی، این ابزارها در شناسایی دقیق‌تر نام مکان از کارآمدی کافی برخوردار نیستند. دلیل این امر را شاید بتوان در نوع نام‌گذاری مکان‌ها یافت. بیشتر نام‌های منتسب به مکان‌های شهری به گونه‌ای است که ابزارهای تشخیص واحد اسمی[۳] آن‌ها را غیرمکانی شناسایی می‌کند. برای مثال، جاینام‌های « هفده شهریور»، «امام خمینی» و «سازمان برنامه و بودجه» بدون آنکه عنوان خیابان پیش از آن‌ها بیاید، به ترتیب به عنوان واحد اسمی زمانی، شخصیت و سازمان شناسایی خواهند شد. لذا استفاده از روش پیشنهادی Hu و همکاران [۱۴] که مبتنی‌بر استفاده از ابزارهای تشخیص واحد اسمی بوده است و به خاطر استفاده از معیارهای زبان‌شناختی از حجم پردازش‌های مکانی کاسته می‌شود، برای زبان فارسی (منطقه مطالعاتی ایران) عملکرد قابل قبولی را ارائه نخواهد داد.

در روشی که McKenzie و همکاران ارائه کرده‌اند [۳۶]، تنها بر اساس توزیع مکانی نقاط هر ان-گرم جاینام‌های احتمالی شناسایی می‌شوند. در مقایسه با پژوهش آن‌ها که منطقه مورد مطالعه سه شهر واشنگتن، سیاتل و مونترال در نظر گرفته شده است، در این تحقیق

---

۱ Hadoop

۲ N-gram، توالی پیوسته‌ی n بخشی از یک نمونه متن یا کلام است.

۳ Named Entity Recognition

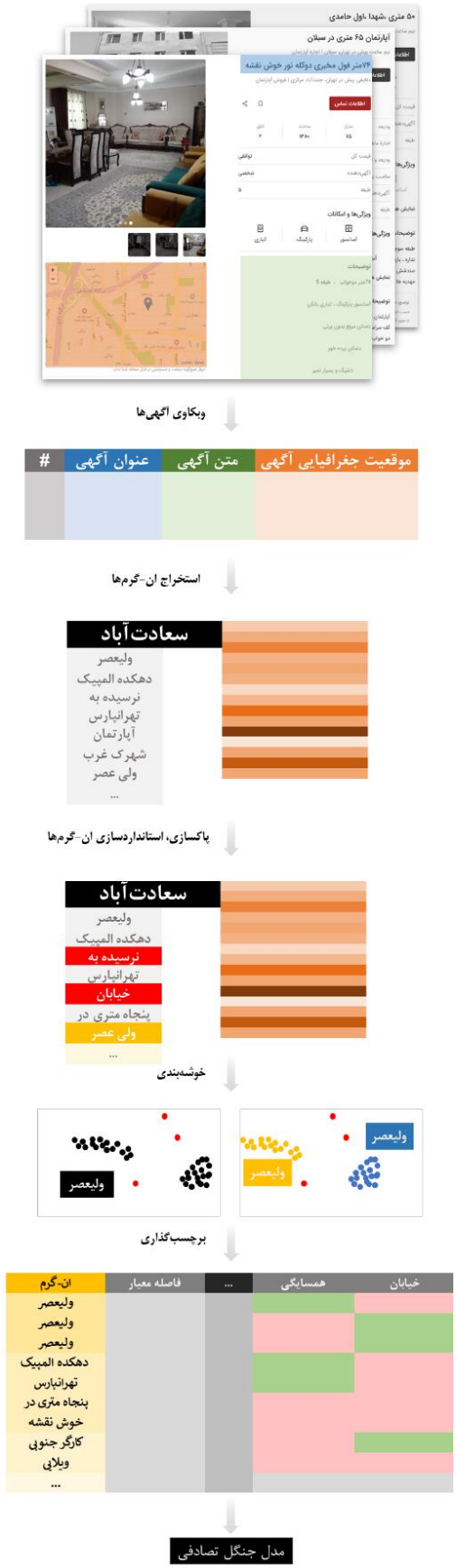
مناطق مورد مطالعه چهار شهر پرجمعیت ایران انتخاب شده است. تفاوت‌های اساسی در زبان، جمعیت، نوع نامگذاری مکان‌ها و نیز نوع آدرس‌دهی در شهرهای انتخاب شده این دو پژوهش مستلزم اعمال تغییراتی در فرآیند ایجاد مدل دسته‌بندی بوده است. علاوه‌بر این، در کنار شناسایی همسایگی‌ها، شناسایی خیابان‌ها در نظر گرفته شد که چالش‌های جدیدی را به همراه داشت.

فرآیند کار از وبکاوی آگهی‌های املاک آغاز می‌شود؛ از آگهی‌های وبکاوی‌شده، ان-گرم‌ها استخراج می‌شوند؛ این ان-گرم‌ها ابتدا پاکسازی شده و سپس استانداردسازی می‌شوند؛ با در نظر گرفتن این سناریو که ممکن است چندین جاینام در موقعیت‌های متفاوتی از شهر نام یکسانی داشته باشند، خوشه‌بندی انجام می‌شود و چنانچه خوشه دارای حداقل نقاط تعیین‌شده باشد به عنوان ان-گرمی مستقل در نظر گرفته می‌شود؛ بر اساس توزیع نقاط بر روی نقشه مبنا برچسب‌گذاری انجام می‌شود؛ برای هر ان-گرم، معیارهای مکانی محاسبه می‌شود؛ مدل تشخیص نوع مکان (همسایگی یا خیابان) ساخته شده و پس از ارزیابی‌های اولیه، بر روی داده سایر شهرها آزموده می‌شود.

۳-۱- داده

در این پژوهش چهار کلانشهر تهران، مشهد، اصفهان و شیراز به عنوان مناطق مورد مطالعه انتخاب شد. از میان چند وبگاه انتشار آگهی املاک، پس از بررسی‌های به عمل آمده، با توجه به تعداد آگهی‌ها و پوشش هر چهار شهر، تنها وبگاه دیوار[۱] انتخاب شد و وبکاوی به مدت یک ماه (ماه ژانویه ۲۰۲۰) انجام گرفت.

در فرآیند وبکاوی، تنها آگهی‌های دارای برچسب مکانی لحاظ شده و به ازای هر آگهی، عنوان آگهی، متن آگهی و طول و عرض جغرافیایی برچسب‌شده به آگهی ذخیره می‌شد. پس از پاکسازی آگهی‌های با مختصات جغرافیایی تکراری، براساس مرزبندی‌های سیاسی، نقاط خارج از محدوده مدنظر حذف شدند.



شکل ۱- روندنما مراحل تهیه و آماده‌سازی داده برای مدل جنگل تصادفی همراه با نمونه‌ای از ان-گرم‌ها

[۱] https://divar.ir

در ادامه، از روی عنوان و متن آگهی، پس از اعمال برخی تصحیحات و نرمال‌سازی، با استفاده از کتابخانه Hazm، توکن‌سازی[1] واژه‌ها انجام شده و توکن‌های یکتایی (مثل «خیابان»، «آزادی» و «گلستان»)، دوتایی (مثل «خیابان بهار»)، سه‌تایی (مثل «شهرک شهید کشوری»)، چهارتایی و پنج‌تایی استخراج شدند. در حالت کلی، هر کدام از این توکن‌های چندتایی تحت عنوان ان-گرم شناخته می‌شوند. به ازای هر ان-گرم، مختصات جغرافیایی تمامی آگهی‌های شامل آن ان-گرم به مجموعه موقعیت‌های جغرافیایی آن ان-گرم اضافه شد. در جدول ۱، تعداد آگهی‌ها و ان-گرم‌های استخراج‌شده از این آگهی‌ها برای هر چهار شهر ارائه شده است.

جدول ۱- تعداد آگهی‌ها و ان-گرم‌های استخراج‌شده

نام شهر	تعداد آگهی‌ها	تعداد ان-گرم‌های استخراج‌شده
تهران	۲۲۳۵۸	۱۱۹۶۷۴۲
مشهد	۱۰۲۸۳	۶۳۶۷۳۲
اصفهان	۷۷۶۸	۴۸۲۱۵۶
شیراز	۷۷۱۲	۳۹۹۱۷۹

پاکسازی ان-گرم‌ها از دو جنبه انجام گرفت. از جنبه آماری، ان-گرم‌های با فراوانی بیشتر از سه برابر انحراف معیار و نیز ان-گرم‌های با فراوانی کمتر از حدآستانه (۱۰ نقطه) تعریف شده حذف شدند. از جنبه غیرآماری، ان-گرم‌های با کمتر از سه کاراکتر و ان-گرم‌هایی که تنها شامل عدد بودند، حذف شدند.

در مرحله استانداردسازی ان-گرم‌ها، با توجه به شیوه نگارش عبارات در آگهی‌ها و سایر مسائل، عبارات و واژه‌ها به شیوه متداول‌شان تغییر یافت. از جمله این تغییرات می‌توان به این موارد اشاره کرد: تغییر حروف اختصاری به واژه مدنظر (برای مثال، تغییر «خ» به «خیابان»)؛ استانداردسازی کاراکترها (تغییر «حسن اباد» به «حسن آباد») و حذف فاصله بین اجزای یک واژه (تغییر «ولی عصر» به «ولیعصر»). در اعمال همه تغییرات، چنانچه ان-گرم جدید قبلاً در مجموعه وجود داشت، مختصات جغرافیایی این دو تلفیق می‌شد.

۳-۲- خوشه‌بندی

خوشه‌بندی، دسته‌بندی مجموعه داده به زیرمجموعه‌هایی تحت عنوان خوشه است به نحوی که اجزای هر زیرمجموعه نسبت به هم شباهت بیشتری در مقایسه با اجزای سایر خوشه‌ها داشته باشد [۳۷].

در بین مجموعه نقاط ان-گرم‌های مکانی، نقاطی بودند که موقعیتشان بسیار دورتر از محدوده جغرافیایی اصلی مکان بود. برای مثال، فرض کنید در میان مجموعه نقاط ان-گرم «پیروزی» که به محله پیروزی واقع در شرق تهران اشاره دارد، سه نقطه با موقعیتی در غرب شهر بوده باشد. برخی از این نقاط ناشی از به اشتباه ثبت کردن مکان آگهی و برخی مربوط به مکان‌هایی با اهمیت بسیار پایین (تکرار بسیار کم در آگهی‌ها مثلاً یک کوچه) و البته دورتر از مکان شناخته‌شده‌تر بوده‌اند. این نقاط به عنوان نقاط پرت تحت خوشه‌بندی DBSCAN از مجموعه نقاط ان-گرم مربوطه حذف شدند [۳۸].

با توجه به اینکه برخی از نام‌ها ممکن است به چند مکان مختلف اشاره داشته باشند و از طرفی در این پژوهش علاوه بر شناسایی همسایگی‌ها، شناسایی خیابان‌ها نیز مدنظر بود، لازم بود مجموعه نقاط آن‌ها از هم تفکیک شود. از این رو، علاوه بر حذف نقاط پرت، مجموعه نقاط هر ان-گرم، خوشه بندی شده و چنانچه تعداد نقاط خوشه از حد کمینه تعریف شده (۱۰ نقطه) بیشتر بود، به عنوان ان-گرمی جدا ذخیره شد.

۳-۳- برچسب گذاری

ابتدا ان-گرم‌ها به لحاظ زبان‌شناختی بررسی شده و در مرحله بعد براساس نمایش نقاط هر ان-گرم بر روی نقشه‌مبنای OpenStreetMap، نوع عارضه جغرافیایی شناسایی و برچسب‌گذاری شد.

برچسب‌گذاری ان-گرم‌ها با دو چالش مهم همراه بود: ۱- تفکیک خیابان از همسایگی و ۲- چندبخشی بودن جاینام‌ها.

۳-۳-۱- تفکیک خیابان از همسایگی

بخش قابل توجهی از نام خیابان‌ها تحت تاثیر نام محله‌ای در همان منطقه بوده است. یا بالعکس، نام محله یا شهرک برگرفته از نام خیابان اصلی‌ای است که در آن محدوده بوده است. ان-گرم‌هایی مثل «حسین‌آباد» اصفهان، «مطهری» مشهد، «سفیر» شیراز و «سهروردی» تهران هم می‌تواند نام همسایگی و هم نام خیابان بوده باشد.

[1] Tokenization

در حالت خاص، ممکن است خیابانی که ارتباط اصلی یک همسایگی را به سایر نقاط شهر برقرار می‌کند، با آن همسایگی همنام باشد. برای مثال در شکل ۲، بخشی از نقاط ان-گرم «کوثر» شامل نقاط بلوار کوثر و بخشی مربوط به شهرک کوثر است. در این حالت، بخشی از نقاط تحت تاثیر خیابان دارای توزیعی خطی و بخش دیگر تحت تاثیر شهرک یا محله‌ای به همان نام، توزیعی غیر خطی خواهد داشت.

در برخی موارد، همسایگی همنام ان-گرم خیابان نیست ولی چون بخشی از خیابان‌های فرعی یا کوچه‌های متصل به خیابان اصلی از نام خیابان اصلی گرفته می‌شوند، توزیع نقاط برای ان-گرم خیابان از حالت خطی خارج می‌شود. برای مثال، در شکل ۳، بخشی از نقاط ان-گرم «نماز» مربوط به بلوار نماز می‌شود و بخش دیگر مربوط به کوچه‌هایی است که در محله آبادگران بوده و در اطراف این بلوار قرار دارند.

شکل ۲- توزیع نقاط ان-گرم «کوثر» - شهر مشهد

شکل ۳- توزیع نقاط ان-گرم «نماز» - شهر مشهد

صرف نظر از شکل هندسی معمول برای همسایگی و خیابان، در برچسب‌گذاری مواردی از جمله سطح تحت پوشش نقاط و نیز تراکم آن‌ها مدنظر بوده است. اگرچه سعی بر این بوده تا ان-گرم تنها در یک دسته برچسب‌گذاری شود، ولی محدودیتی برای برچسب گذاری در دو دسته خیابان و همسایگی اعمال نشد.

۳-۳-۲- چندبخشی بودن جاینام‌ها

چالش چندبخشی بودن جاینام (بدون تعیین نوع مکان)، به این معنا است که بخش اول، دوم و ... جاینام جداگانه شامل همان نقاطی باشد که خود جاینام داراست. برای مثال، جاینام «گلدشت حافظ» در شهر شیراز به عنوان یک جاینام از نوع همسایگی از دو بخش «گلدشت» و «حافظ» ساخته شده است. مسلماً نه «گلدشت» و نه «حافظ» به تنهایی نمی‌توانند به عنوان جاینام در نظر گرفته شوند (حداقل در کاربرد رسمی). ولی مجموعه نقاط مربوط به جاینام «گلدشت حافظ» همان مجموعه نقاط ان-گرم‌های «گلدشت» و «حافظ» است. در نتیجه مقادیر معیارهای مکانی محاسبه شده برای هر سه با هم برابر خواهد بود. روندی که در برچسب‌گذاری اینگونه ان-گرم‌ها لحاظ شد تنها در نظر گرفتن ان-گرم جاینام کامل و حذف سایر ان-گرم‌های با مختصات یکسان بود.

علاوه‌بر این، ان-گرم‌هایی که نوع مکان را نیز شامل بودند (مثل «شهرک ولیعصر») در کنار حالت بدون نوع مکان (مثل «ولیعصر») به عنوان همسایگی یا خیابان برچسب‌گذاری شدند. در خصوص ان-گرم‌های مربوط به شخصیت‌ها نظیر «حکیم نظامی» و «شهید وفایی»، با توجه به کاربرد و نوع عنوان منسوب به شخصیت در برخی موارد تنها حالت همراه با عنوان منسوب و در برخی موارد هر دو حالت برچسب‌گذاری شد. در برچسب‌گذاری ان-گرم‌های مربوط به رویدادهای تاریخی مثل «هفده شهریور» یا «خیابان بیست و دو بهمن» نیز تنها شکل کامل آن‌ها به عنوان ان-گرم همسایگی یا خیابان در نظر گرفته شد و حالات دیگر مثل «خیابان بیست و دو» یا «شهریور» به خاطر داشتن مجموعه نقاط یکسان حذف شدند.

جدول ۲- تعداد ان-گرم‌های با برچسب خیابان، همسایگی و کل ان-گرم‌ها برای هر چهار شهر

شهر	خیابان	همسایگی	کل
تهران	۱۶۸	۳۸۶	۵۸۱۲
مشهد	۹۲	۱۴۸	۲۵۶۱
اصفهان	۱۶۱	۱۹۳	۱۸۰۵
شیراز	۹۲	۱۱۴	۱۵۵۲

۳-۴- معیارهای مکانی

در این تحقیق فرض بر این بوده است که دسته‌بندی عارضه جغرافیایی می‌تواند بر اساس معیارهایی مشخص شود. برای شناسایی همسایگی‌ها و خیابان‌ها، ۲۴ معیار مکانی که در پژوهش McKenzie و همکاران اتخاذ شده بود [۳۶]، در این مطالعه هم به عنوان شاخصه‌های تعیین کننده نوع عارضه جغرافیایی در نظر گرفته شد. با توجه به شکل ۴، این معیارها اگرچه شاید در برخی موارد دارای همبستگی بالایی باشند ولی همانطور که در شکل‌های ۵، ۶، ۷ و ۸ نشان داده شده است، میزان اهمیت هر کدام از آن‌ها برای تفکیک نوع عوارض برای هر شهر متفاوت است.

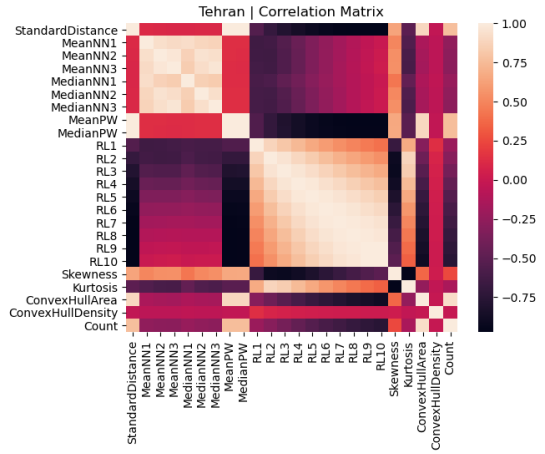
برای پراکندگی مکانی[۱]، ۹ معیار محاسبه شده به ازای هر ان-گرم عبارت‌اند از: فاصله معیار[۲]؛ متوسط فاصله تا نزدیک‌ترین همسایگی برای اولین، دومین و سومین همسایگی، بصورت میانگین و میانه؛ متوسط فاصله زوج نقاط از هم، بصورت میانگین و میانه.

برای همگنی مکانی[۳]، مقادیر تابع Ripley's L در محدوده ۰ تا ۵ کیلومتر در ۱۰ زیربازه ۵۰۰ متری محاسبه و مقادیر مربوط به هر بازه به عنوان معیاری جدا به ازای هر ان-گرم ذخیره شد. علاوه‌بر این ۱۰ معیار، کشیدگی و چولگی تابع Ripley's L نیز محاسبه گردید.

همچنین، تعداد نقاط، مساحت محدوده تحت پوشش و تراکم نقاط به عنوان سه معیار آخر محاسبه شد. از این رو، پوش محدب برای مجموعه نقاط هر ان-گرم ساخته شده و مساحت پوش محدب محاسبه شد. تراکم نقاط نیز با تقسیم تعداد نقاط بر مساحت پوش محدب بدست آمد.

با توجه به شکل‌های ۵، ۶، ۷ و ۸، در شناسایی خیابان، در همه‌ی شهرها معیار تراکم نقاط تاثیرگذارترین معیار بوده است. پس از آن، معیارهای مربوط به متوسط فاصله تا نزدیک‌ترین همسایگی از اهمیت بالاتری برخوردار بوده‌اند. در حالت کلی، میزان اهمیت معیارهای کشیدگی و چولگی بیشتر از ده معیار مربوط به مقدار خود تابع Ripley's L بوده است؛ اگر چه در مدل تولید شده برخی شهرها، مقدار تابع برای فواصل پایین‌تر (RL2 و RL3) نیز از معیارهای مهم بوده است. معیار تعداد نقاط و فاصله استاندارد در اکثر شهرها جزء معیارهای کم اهمیت برای عمکرد مدل در شناسایی ان-گرم‌های خیابان بوده است.

در شناسایی همسایگی، معیار تراکم همچنان از معیارهای با اهمیت بالا بوده است. از مجموعه معیارهای همگنی مکانی، معیارهای چولگی و کشیدگی دارای اهمیت بالایی بوده‌اند اگرچه نقش مقادیر تابع در عملکرد مدل نسبت به شناسایی خیابان تا حدودی افزایش یافته است.



شکل ۴- میزان همبستگی ۲۴ معیار برای مدل آموزش‌یافته براساس داده شهر تهران

۳-۵- مدلسازی

روش یادگیری جمعی جنگل‌های تصادفی، روشی است که با ایجاد تعداد زیادی درخت تصمیم، مسئله طبقه‌بندی یا برازش حل می‌شود [۳۹]. سهولت در استفاده از این روش با توجه به این که به پارامترهای کمی برای تنظیم مدل نیاز است و دقت بالا باعث شده است تا در مسائل مختلفی از آن استفاده شود [۴۰-۴۲]. علاوه‌بر این، روش جنگل‌های تصادفی به خاطر قابلیت تعامل با اندازه نمونه کوچک و فضاهای ویژگی چند بعدی شناخته می‌شود [۴۳].

با توجه به نتایج حاصل از بکارگیری روش جنگل‌های تصادفی نسبت به استفاده جداگانه از هر کدام از معیارها در پژوهش McKenzie و همکاران [۳۶]، در این مطالعه نیز روش جنگل‌های تصادفی برای طبقه‌بندی اتخاذ شد. ابتدا، حد آستانه بهینه برای احتمال پیش‌بینی تعیین شده و سپس، مدل ساخته‌شده براساس حد آستانه بهینه برای هر چهار شهر بر روی سایر شهرها آزموده شد.

برای مشخص کردن بهترین حد آستانه، در ۱۰ حلقه تکرار، نتیجه حاصل از میانگین ۱۰۰ مدل به عنوان نتیجه نهایی هر حدآستانه (از ۰/۱ تا ۱) ذخیره شد. در هر بار که

۱ Spatial Dispersion
۲ Standard Distance
۳ Spatial Homogeneity

مدلی ساخته می‌شد، بطور تصادفی ۷۰٪ از ان-گرم‌ها به عنوان داده آموزشی انتخاب شده و بر روی ۳۰٪ باقیمانده آزموده می‌شد. براساس مقدار F_score میانگین ۱۰۰ تکرار برای هر حدآستانه، حدآستانه بهینه برای آزمون مدل ساخته‌شده هر شهر بر روی داده سایر شهرها انتخاب شد.

با توجه به طبیعت مسئله، سهم ان-گرم‌های با برچسب همسایگی یا خیابان از کل ان-گرم‌ها پایین بود. تعداد بسیار کم ان-گرم‌های با مقدار مثبت (هم برای برچسب همسایگی و هم خیابان) باعث می‌شود در مرحله آموزش مدل، مدل بیشتر به ان-گرم‌های با مقدار منفی گرایش داشته باشد. به همین دلیل، پس از هر بار تفکیک مجموعه ان-گرم‌ها به آموزشی و ارزیابی، روی مجموعه ان-گرم‌های آموزشی نمونه‌برداری دوباره انجام گرفت و سهم ان-گرم‌های با برچسب مثبت افزایش یافته و برابر با سهم ان-گرم‌های منفی شد.

علاوه‌بر F_score که در بالا گفته شد، سه معیار صحت (Accuracy)، دقت (Precision) و بازیابی (Recall) نیز محاسبه شدند. با این وجود، در ارزیابی عملکرد مدل‌ها، با توجه به این که معیار صحت نمی‌تواند به عنوان معیاری قابل اعتماد باشد، تنها از سه معیار دقت، بازیابی و F_score استفاده شد.

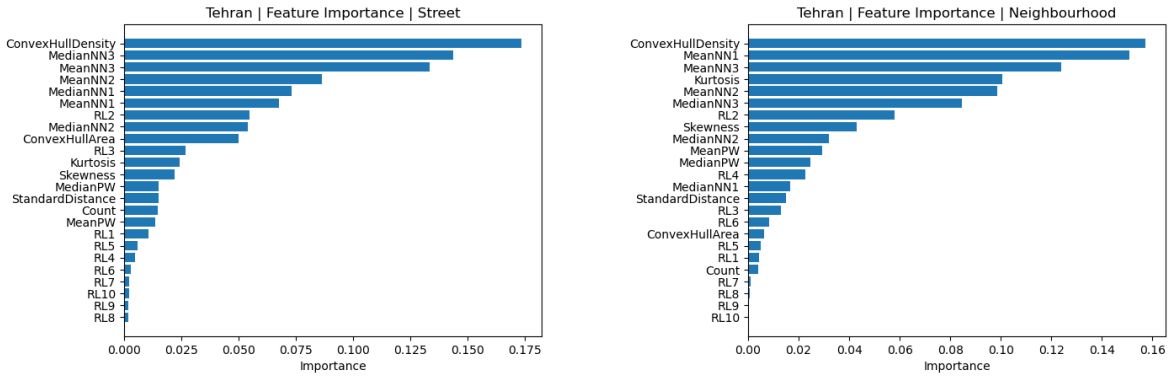
موارد برآوردشده بر اساس مدل از چهار حالت خارج نیست که این چهار حالت عبارت‌اند از: ۱- مواردی که مطابق آنچه که به عنوان خیابان یا همسایگی برچسب‌گذاری شده بودند، شناسایی شده‌اند (T_p)؛ ۲- مواردی که به اشتباه به عنوان همسایگی یا خیابان شناسایی شده‌اند (F_p)؛ ۳- مواردی که برخلاف برچسبی که زده شده بود به اشتباه به عنوان همسایگی یا خیابان شناسایی نشده‌اند (F_N) و ۴- مواردی که به درستی به عنوان همسایگی یا خیابان شناسایی نشده‌اند (T_n). در ادامه فرمول هر چهار معیار بر اساس حالات گفته شده، آورده شده است:

$$\text{Accuracy} = \frac{T_P+T_N}{T_P+T_N+F_P+F_N} \quad (۱)$$

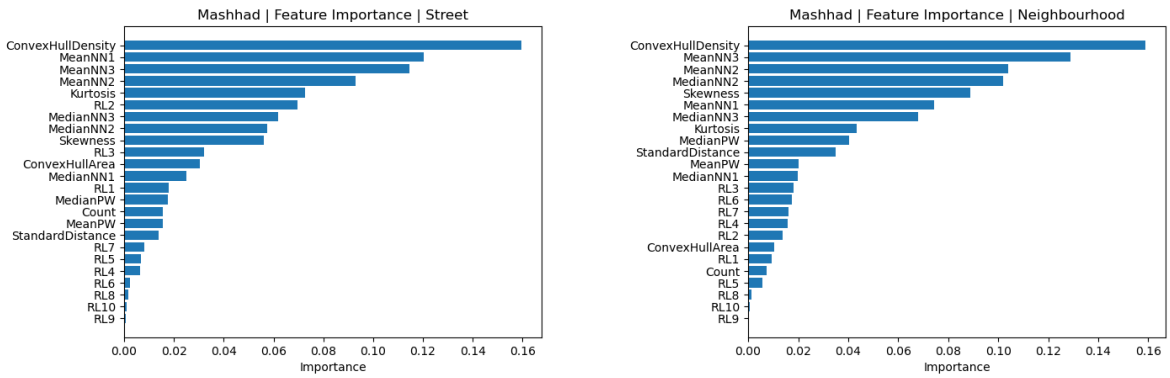
$$\text{Precision} = \frac{T_P}{T_P+F_P} \quad (۲)$$

$$\text{Recall} = \frac{T_P}{T_P+F_N} \quad (۳)$$

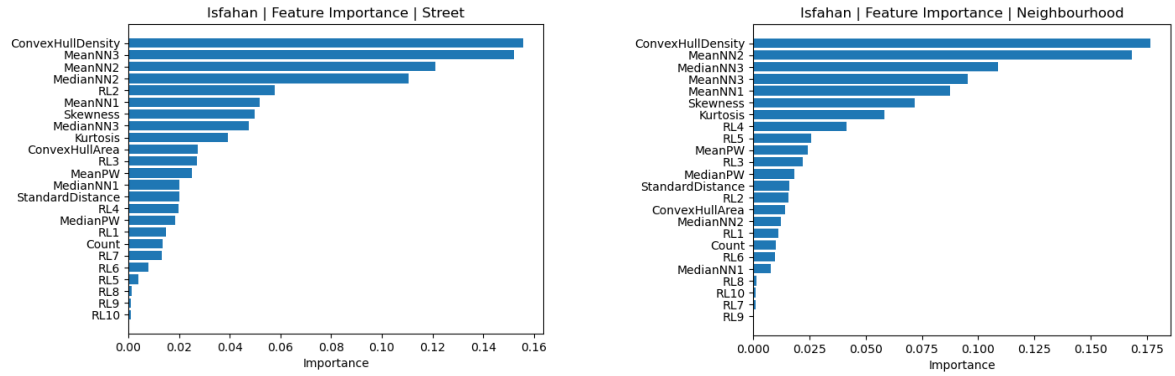
$$F_{score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{Precision + Recall} \quad (۴)$$



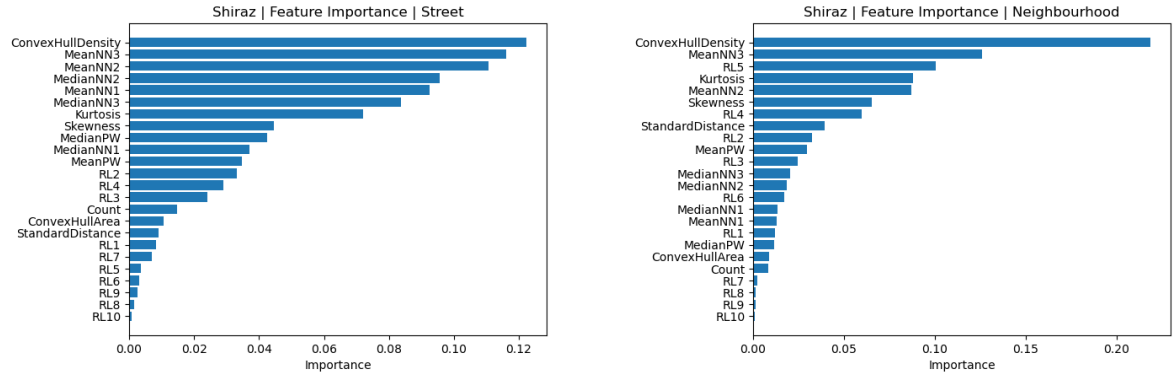
شکل ۵- نمودارهای میزان اهمیت ۲۴ معیار مکانی برای شناسایی همسایگی و خیابان - شهر تهران



شکل ۶- نمودارهای میزان اهمیت ۲۴ معیار مکانی برای شناسایی همسایگی و خیابان - شهر مشهد



شکل ۷- نمودارهای میزان اهمیت ۲۴ معیار مکانی برای شناسایی همسایگی و خیابان - شهر اصفهان



شکل ۸- نمودارهای میزان اهمیت ۲۴ معیار مکانی برای شناسایی همسایگی و خیابان - شهر شیراز

۴- نتایج

در این بخش، نتایج آزمون مدل‌های ساخته‌شده هر شهر بر مبنای ۷۰٪ از کل داده و آزمون بر روی ۳۰٪ باقیمانده از داده همان شهر و برروی کل داده سایر شهرها ارائه شده است. بر اساس سه معیار عنوان شده در بخش قبل، عملکرد مدل‌های تولیدشده در شناسایی همسایگی و خیابان مورد ارزیابی قرار گرفت. همانطور که در جداول ۳، ۴ و ۵ ارائه شده است، اگرچه مدل‌های تولیدشده از میزان بازیابی بالایی برخورد بوده‌اند ولی مقدار معیار دقت باعث کاهش مقدار *F_score* شده است. دلایل این امر را بایستی از دو جنبه کمی و کیفی مدنظر داشت.

از جنبه کمی، سهم کم ان-گرم‌های با برچسب خیابان یا همسایگی از کل ان-گرم‌ها باعث می‌شود معیار دقت محاسبه‌شده برای مدل کم باشد. اگرچه با انجام نمونه‌برداری و رفع عدم توازن، احتمال مشارکت ان-گرم‌های با برچسب مثبت در آموزش مدل افزایش می‌یابد ولی همچنان به خاطر محدود بودن دامنه تغییرات معیارها، مدل تنها بر اساس شکل خاصی از ان-گرم‌های با برچسب مثبت آموزش می‌یابد. نتیجه این می‌شود که مدل آموزش‌یافته یک شهر، حتی در شناسایی ان-گرم‌های خود آن شهر عملکرد نه چندان خوبی را از نظر معیار دقت ارائه کند.

از جنبه کیفی، پایین بودن مقدار دقت، به ویژه در شناسایی خیابان، بیشتر تحت تاثیر مجموعه ان-گرم‌های با برچسب مثبت است. به بیانی دیگر، سیاست‌های اتخاذشده در برچسب‌گذاری ان-گرم‌ها هم در آزمون مدل آموزش‌یافته بر روی داده خود آن شهر و هم روی داده سایر شهرها تاثیر عمده‌ای داشته است. با بررسی تغییرات صورت گرفته در پیش‌بینی یک ان-گرم به عنوان خیابان یا همسایگی نسبت به آنچه که برچسب‌گذاری شده بود، می‌توان گفت بخش زیادی از کاهش دقت ناشی از شناسایی یک همسایگی به عنوان یک خیابان یا بالعکس بوده است.

بخشی از کاهش مقدار دقت، ان-گرم‌های غیر مکانی هستند که به اشتباه به عنوان همسایگی یا خیابان پیش‌بینی می‌شوند. بخشی از این گونه ان-گرم‌ها در واقع بر گرفته از ادبیات خاص هر منطقه بیشترین هم‌رخدادی را با ان-گرم خیابان یا همسایگی مربوط به آن منطقه دارند. از این رو، رفتاری مشابه ان-گرم خیابان یا همسایگی را ایفا می‌کنند. آنچه که این مسئله را تقویت

می‌کند انحصار انتشار آگهی‌ها توسط مشاورین املاک برای منطقه‌ای خاص از شهر است. در اکثر آگهی‌هایی که توسط مشاورین املاک منتشر می‌شود، معمولاً مجموعه عبارات مشترکی چه برای معرفی ملک و چه برای تبلیغات نوشته می‌شود.

تعداد کمی از ان-گرم‌های با برچسب منفی نیز تحت تاثیر خوشه‌بندی و تشکیل خوشه‌هایی منطبق بر خوشه‌های خیابان و همسایگی، به اشتباه به عنوان ان-گرمی با برچسب مثبت پیش‌بینی می‌شوند. برای مثال، فرض کنید ان-گرمی مثل «فروش آپارتمان» که مجموعه نقاطش در کل شهر اصفهان توزیع می‌یابد، خوشه‌بندی شود. پس از خوشه‌بندی، یک خوشه بخش اصلی شهر را پوشش می‌دهد و خوشه دیگر سپاهانشهر را که از بخش اصلی شهر دورتر است. در نتیجه، مقادیر محاسبه شده معیارها برای ان-گرم «فروش آپارتمان» خوشه دوم بسیار شبیه مقادیری خواهد بود که برای ان-گرم «سپاهانشهر» محاسبه شده است.

آنچه که معیار بازیابی را در آزمون مدل روی داده شهری دیگر بیشتر تحت تاثیر قرار داده است را شاید بتوان تفاوت‌های عمده در الگوهای شهرسازی هر شهر دانست. برای مثال، در مجموعه ان-گرم‌های برچسب‌گذاری شده به عنوان خیابان شهر مشهد، خیابان‌هایی هستند که امتداد ثابتی ندارند و در عین حال توزیع نقاط به گونه‌ای است که باعث می‌شود چنین ان-گرم‌هایی در آزمون مدل شهری دیگر، به احتمال زیاد به عنوان همسایگی پیش‌بینی شوند. ان-گرم‌هایی مثل «اتوبان امام علی» و «خیابان کارگر» تهران نیز به این خاطر که ان-گرم‌هایی با شرایط مکانی آن‌ها در سایر شهرها نیست یا بسیار کم است، توسط مدل آموزش‌یافته سایر شهرها به عنوان یک ان-گرم غیر مکانی پیش‌بینی می‌شوند.

باتوجه به جدول ۴، در شناسایی خیابان، مدل تولیدشده براساس داده شهر تهران در آزمون بر روی داده سایر شهرها بیشترین دقت را داشته است. در شناسایی همسایگی نیز مدل تولیدشده بر اساس داده شهر مشهد بالاترین دقت را در شناسایی همسایگی‌ها ارائه کرده است. در آزمون مدل آموزش‌یافته بر مبنای ۷۰٪ داده یک شهر و آزمون بر روی ۳۰٪ باقیمانده از داده همان شهر، در شناسایی خیابان و همسایگی به ترتیب، مدل‌های شهر شیراز و تهران بالاترین دقت را داشته اند. از طرفی، داده شهر اصفهان در شناسایی خیابان و داده شهر تهران در شناسایی همسایگی توسط مدل سایر شهرها با دقت بالاتری مورد پیش‌بینی قرار گرفته‌اند.

مطابق جدول ۵، مدل تولیدشده برمبنای داده شهر اصفهان در شناسایی خیابان و مدل شهر شیراز در شناسایی همسایگی بالاترین بازیابی را در آزمون روی داده سایر شهرها داشته اند. در آزمون مدل یک شهر بر روی داده همان شهر نیز در شناسایی خیابان، مدل شهر مشهد و در شناسایی همسایگی، مدل شهر تهران بالاترین بازیابی را داشته اند. همچنین، داده شهر تهران در شناسایی خیابان و همسایگی توسط مدل‌های سایر شهرها با بازیابی بالاتری در مقایسه با داده بقیه شهرها مورد پیش‌بینی قرار گرفته است.

جدول ۳- مقادیر معیار F_score برای مدل‌های آموزش یافته براساس ۷۰٪ داده یک شهر و آزمون روی ۳۰٪ باقیمانده و نیز آزمون روی سایر شهرها

		شهر در نظر گرفته شده برای آزمون							
		تهران		مشهد		اصفهان		شیراز	
		خیابان	همسایگی	خیابان	همسایگی	خیابان	همسایگی	خیابان	همسایگی
شهر در نظر گرفته شده برای آموزش مدل	تهران	۰/۵۸۴	۰/۸۴۷	۰/۶۱۳	۰/۷۴۳	۰/۵۹۸	۰/۷۷۲	۰/۵۲۰	۰/۷۱۷
	مشهد	۰/۵۰۲	۰/۸۱۱	۰/۶۴۸	۰/۷۷۵	۰/۶۷۰	۰/۷۶۸	۰/۶۰۱	۰/۶۹۱
	اصفهان	۰/۴۲۷	۰/۸۱۶	۰/۵۵۰	۰/۶۹۳	۰/۶۸۲	۰/۸۰۳	۰/۵۹۶	۰/۶۹۲
	شیراز	۰/۵۰۲	۰/۸۰۵	۰/۶۲۷	۰/۷۴۰	۰/۶۱۴	۰/۷۵۲	۰/۷۲۵	۰/۶۰۰

جدول ۴- مقادیر معیار دقت برای مدل‌های آموزش یافته براساس ۷۰٪ داده یک شهر و آزمون روی ۳۰٪ باقیمانده و نیز آزمون روی سایر شهرها

		شهر در نظر گرفته شده برای آزمون							
		تهران		مشهد		اصفهان		شیراز	
		خیابان	همسایگی	خیابان	همسایگی	خیابان	همسایگی	خیابان	همسایگی
شهر درنظر گرفته شده برای آموزش مدل	تهران	۰/۴۶۰	۰/۷۶۳	۰/۶۰۶	۰/۶۶۰	۰/۵۹۵	۰/۷۰۰	۰/۴۹۰	۰/۶۰۶
	مشهد	۰/۳۵۵	۰/۷۲۳	۰/۵۰۰	۰/۶۹۱	۰/۵۷۹	۰/۶۷۴	۰/۴۸۳	۰/۵۹۰
	اصفهان	۰/۲۸۰	۰/۷۲۲	۰/۳۹۲	۰/۵۷۶	۰/۵۵۰	۰/۷۴۷	۰/۴۴۲	۰/۵۷۱
	شیراز	۰/۳۷۴	۰/۶۹۸	۰/۵۴۴	۰/۶۳۶	۰/۵۶۵	۰/۶۴۰	۰/۶۷۶	۰/۴۴۴

جدول ۵- مقادیر معیار بازیابی برای مدل‌های آموزش یافته براساس ۷۰٪ داده یک شهر و آزمون روی ۳۰٪ باقیمانده و نیز آزمون روی سایر شهرها

		شهر در نظر گرفته شده برای آزمون							
		تهران		مشهد		اصفهان		شیراز	
		خیابان	همسایگی	خیابان	همسایگی	خیابان	همسایگی	خیابان	همسایگی
شهر درنظر گرفته شده برای آموزش مدل	تهران	۰/۸۰۰	۰/۹۵۲	۰/۶۱۹	۰/۸۵۱	۰/۶۰۲	۰/۸۶۰	۰/۵۵۴	۰/۸۷۷
	مشهد	۰/۸۵۷	۰/۹۲۵	۰/۹۲۰	۰/۸۸۴	۰/۷۹۵	۰/۸۹۱	۰/۷۹۳	۰/۸۳۳
	اصفهان	۰/۹۰۵	۰/۹۳۸	۰/۹۲۴	۰/۸۷۲	۰/۸۹۸	۰/۸۶۸	۰/۹۱۳	۰/۸۷۷
	شیراز	۰/۷۶۲	۰/۹۵۱	۰/۷۳۹	۰/۸۸۵	۰/۶۷۱	۰/۹۱۲	۰/۷۸۱	۰/۹۲۳

۵- نتیجه‌گیری و پیشنهادات

جاینامه‌ها به عنوان بخش مهمی از فرآیند بازیابی اطلاعات مکانی، برای کارآمدی در قبال کاربردهای روزافزون این حوزه، بایستی غنی‌سازی شوند. مهم‌ترین جنبه غنی‌سازی جاینامه گردآوری نام‌های جغرافیایی جدید و به‌هنگام‌سازی جاینامه با این جاینام‌ها است. آگهی‌های آنلاین املاک از این نظر که توسط خود مردم منتشر می‌شوند و از دیگر سو، چون شامل جدیدترین جاینام‌ها هستند، می‌توانند به عنوان منبعی برای گردآوری نام‌های جغرافیایی در نظر گرفته شود. با این وجود، متون آگهی‌ها بسیار متفاوت از متون غیرساختار یافته‌ای است که در منابعی مثل اخبار می‌توان یافت. این امر برای زبان فارسی که ابزارهای تشخیص موجودیت اسمی توسعه یافته برای آن به بلوغ کافی نرسیده‌اند، پررنگ‌تر می‌شود. منتشرکنندگان آگهی با فرض این که یک همسایگی یا یک خیابان اصلی توسط خواننده شناخته شده است، تنها نام آن را در آگهی می‌آورند بدون آنکه واژه‌ای در شناساندن مکانی بودن یا نبودن آن و در وهله بعد نوع مکان ذکر شود. این چالش‌ها مستلزم این بودند که روشی متفاوت‌تر از تشخیص موجودیت اسمی برای شناسایی همسایگی‌ها و خیابان‌های اصلی اتخاذ شود.

در این پژوهش، هدف ارائه روشی برای شناسایی جاینام‌های از نوع همسایگی و خیابان با استفاده از آگهی‌های املاک بوده است. آگهی‌های با برچسب مکانی از وبگاه دیوار وبکاوی شده و از روی آگهی‌ها ان-گرم‌ها استخراج شدند. پس از پیش‌پردازش، این ان-گرم‌ها برچسب‌گذاری شدند و در نهایت مدل جنگل‌های تصادفی آموزش‌یافته براساس بخشی از داده بر روی بخش باقیمانده آزموده شد. شناسایی همسایگی در کنار خیابان مستلزم این بود سناریو تکرار یک جاینام برای چند مکان متفاوت درنظر گرفته شود. در این سناریو، مجموعه نقاط هر ان-گرم خوشه‌بندی شده و در نتیجه این خوشه‌بندی، چنانچه خوشه‌ای از تعداد کافی نقاط

برخوردار بود، به عنوان ان-گرمی مستقل در نظر گرفته شد. این امر باعث شد که ان-گرمی مثل «گلستان» برای شهر تهران به جای آن که به خاطر عدم اشاره به یک مکان مشخص حذف شود، در قالب چند ان-گرم مستقل در مجموعه ان-گرم‌ها حضور داشته باشد.

نتایج بدست آمده نشان‌دهنده این است که علیرغم برخی تفاوت‌ها از جمله جمعیت، الگوهای شهرسازی و الگوهای نامگذاری خیابان‌ها و همسایگی‌ها، مدل آموزش‌یافته براساس داده یک شهر، می‌تواند ان-گرم‌های خیابان و همسایگی شهرهای دیگر را نیز به خوبی شناسایی کند. با این حال، با توجه به دلایلی که در بخش ۴ گفته شد، ممکن است ان-گرم‌های برچسب‌گذاری‌شده به عنوان خیابان به عنوان همسایگی پیش‌بینی شوند یا بالعکس، که این امر تحت تاثیر سیاست‌هایی است که در برچسب‌گذاری خیابان و همسایگی اعمال می‌شود. لذا برای پژوهش‌های آتی پیشنهاد می‌شود که در برچسب‌گذاری ان-گرم‌ها جوانب مختلفی مدنظر قرار گیرد.

مسلماً به خاطر در نظر گرفتن بازه زمانی یک ماهه وبکاوی، بخشی از جاینام‌ها شانس حضور در مجموعه ان-گرم‌های با برچسب همسایگی یا خیابان را نداشته‌اند؛ از این رو، در تحقیقات بعدی بهتر است با افزایش دوره زمانی وبکاوی و در نتیجه افزایش ان-گرم‌های همسایگی و خیابان، مدل براساس ان-گرم‌های مکانی بیشتری آموزش یابد. از طرف دیگر، مدل تولیدشده براساس داده یک ماه شاید نتواند ان-گرم‌های استخراج‌شده براساس آگهی‌های ماه‌های دیگر را با توجه به تغییراتی که در مقادیر معیارها ایجاد می‌شود، به خوبی شناسایی کند.

مراجع

[1] Hill, L.L. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. in Research and Advanced Technology for Digital Libraries. 2000. Berlin, Heidelberg: Springer Berlin Heidelberg.

[2] Singh, S.K. and D. Rafiei. Strategies for geographical scoping and improving a gazetteer. in Proceedings of the 2018 World Wide Web Conference. 2018. International World Wide Web Conferences Steering Committee.

[3] Keßler, C., K. Janowicz, and M. Bishr. An agenda for the next generation gazetteer: Geographic information contribution and retrieval. in Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems. 2009. ACM.

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

[5] Di Rocco, L., et al., Sherloc: a knowledge-driven algorithm for geolocating microblog messages at sub-city level. International Journal of Geographical Information Science, 2020: p. 1-32.

[6] Al-Olimat, H.S., et al., Location name extraction from targeted text streams using gazetteer-based statistical language models. arXiv preprint arXiv:1708.03105, 2017.

[7] Javidaneh, A. and F. Karimipour, An Approach for Automatic Matching of Descriptive Addresses. ISSGE, 2020. 9(4): p. 1-17.

[8] Karimzadeh, M., et al., GeoTxt: A scalable geoparsing system for unstructured text geolocation. Transactions in GIS, 2019. 23(1): p. 118-136.

[9] Cura, R., et al., Historical collaborative geocoding. ISPRS International Journal of Geo-Information, 2018. 7(7): p. 262.

[10] Li, D., T.J. Cova, and P.E. Dennison, Using reverse geocoding to identify prominent wildfire evacuation trigger points. Applied geography, 2017. 87: p. 14-27.

[11] Li, H., J. Liu, and X. Zhou, Intelligent map reader: A framework for topographic map understanding with deep learning and gazetteer. IEEE Access, 2018. 6: p. 25363-25376.

[12] Jones, C.B., et al., Modelling vague places with knowledge from the Web. International Journal of Geographical Information Science, 2008. 22(10): p. 1045-1065.

[13] Montello, D.R., et al., Where's downtown?: Behavioral methods for determining referents of vague spatial queries. Spatial Cognition & Computation, 2003. 3(2-3): p. 185-204.

[14] Hu, Y., H. Mao, and G. McKenzie, A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. International Journal of Geographical Information Science, 2019. 33(4): p. 714-738.

[15] F. Goodchild, M. and L. L. Hill, Introduction to digital gazetteer research. Vol. 22. 2008. 1039-1044.

[16] Gelernter, J., et al. Automatic gazetteer enrichment with user-geocoded data. in Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information. 2013. ACM.

[17] Zhang, Y., et al., Extracting geographic features from the Internet: A geographic information mining framework. Knowledge-Based Systems, 2019. 174: p. 57-72.

[18] Ahlers, D. Assessment of the accuracy of GeoNames gazetteer data. in Proceedings of the 7th workshop on geographic information retrieval. 2013.

[19] Acheson, E., S. De Sabbata, and R.S. Purves, A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. Computers, Environment and Urban Systems, 2017. 64: p. 309-320.

[20] Keßler, C., et al. Bottom-up gazetteers: Learning from the implicit semantics of geotags. 2009. Springer

[21] Zhu, R., et al., Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. Transactions in GIS, 2016. 20(3): p. 333-355.

[22] Machado, I.M.R., et al., An ontological gazetteer and its application for place name disambiguation in text. Journal of the Brazilian Computer Society, 2011. 17(4): p. 267-279.

[23] Acheson, E., M. Volpi, and R.S. Purves, Machine learning for cross-gazetteer matching of natural features. International Journal of Geographical Information Science, 2020. 34(4): p. 708-734.

[24] Martins, B., H. Galhardas, and N. Gonçalves. Using Random Forest classifiers to detect duplicate gazetteer records. 2012. IEEE.

[25] Hastings, J. and L. Hill, Treatment of duplicates in the alexandria digital library gazetteer. Proceedings of GeoScience, 2002.

[26] Grossner, K., K. Janowicz, and C. Keßler, Place, period, and setting for linked data gazetteers. Placing names: Enriching and integrating gazetteers, 2016: p. 80-96.

[27] Southall, H., P. Aucott, and M. Stoner, PastPlace linked data historical gazetteer, in Comprehensive Geographic Information Systems. 2017, Elsevier.

[28] Laurini, R., Geographic ontologies, gazetteers and multilingualism. Future Internet, 2015. 7(1): p. 1-23.

[29] de Oliveira, M.G., et al. Leveraging VGI for gazetteer enrichment: A case study for geoparsing twitter messages. 2015. Springer.

[30] Gelernter, J., et al. Automatic gazetteer enrichment with user-geocoded data. in Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information. 2013. ACM.

[31] Popescu, A., G. Grefenstette, and P.A. Moëllic. Gazetiki: automatic creation of a geographical gazetteer. in Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries. 2008. ACM.

[32] Fize, J., G. Shrivastava, and P.A. Ménard. Geodict: an integrated gazetteer. in Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017). 2017.

[33] Twaroch, F.A. and C.B. Jones. A web platform for the evaluation of vernacular place names in automatically constructed gazetteers. in Proceedings of the 6th Workshop on Geographic Information Retrieval. 2010.

[34] Lim, J., et al., Constructing Geographic Dictionary from Streaming Geotagged Tweets. ISPRS International Journal of Geo-Information, 2019. 8(5): p. 216.

[35] Gao, S., et al., Constructing gazetteers from volunteered big geo-data based on Hadoop. Computers, Environment and Urban Systems, 2017. 61: p. 172-186.

[36] McKenzie, G., et al., Identifying urban neighborhood names through user-contributed online property listings. ISPRS International Journal of Geo-Information, 2018. 7(10): p. 388.

[37] Madhulatha, T.S., An overview on clustering methods. arXiv preprint arXiv:1205.1117, 2012.

[38] Schubert, E., et al., DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 2017. 42(3): p. 1-21.

[39] Breiman, L., Random Forests. Machine Learning, 2001. 45(1): p. 5-32.

[40] Felton, B.R., et al., Using random forest classification and nationally available geospatial data to screen for wetlands over large geographic regions. Water, 2019. 11(6): p. 1158.

[41] Chang, S., et al., Mapping the Essential Urban Land Use in Changchun by Applying Random Forest and Multi-Source Geospatial Data. Remote Sensing, 2020. 12(15): p. 2488.

[42] Tehrany, M.S., et al., A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using logitboost machine learning classifier and multi-source geospatial data. Theoretical and Applied Climatology, 2019. 137(1): p. 637-653.

[43] Biau, G. and E. Scornet, A random forest guided tour. Test, 2016. 25(2): p. 197-227.