نشریه علمی - پژوهشی علوم و فنون نقشه برداری، دوره ششم، شماره ۴، اردیبهشت ماه ۱۳۹۶

استخراج نشانه‌ها از اسناد مکانی در موتورهای جستجو

سعید برهانی نژاد [۱]، فرشاد حکیم‌پور*[۲]، احسان حمزه‌ئی[۱]

[۱]دانشجوی کارشناسی ارشد سیستم‌های اطلاعات مکانی - دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی - پردیس دانشکده‌های فنی - دانشگاه تهران
{saeid.borhani, e.hamzei}@ut.ac.ir

[۲]استادیار دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی - پردیس دانشکده‌های فنی - دانشگاه تهران
fhakimpour@ut.ac.ir

(تاریخ دریافت شهریور ۱۳۹۵، تاریخ تصویب بهمن ۱۳۹۵)

چکیده

امروزه امکان دسترسی انتخابی به اطلاعات بر روی وب، از طریق موتورهای جستجو فراهم می‌شود. اما در مواردی که نیاز ما در برگیرنده جستجو در اطلاعات مکانی نیز باشد وظیفه جستجو پیچیده‌تر می‌شود و احتیاج به توانایی‌های خاصی در بخش جستجوگر است. هدف اصلی انجام این پژوهش ایجاد بستری جهت استخراج اطلاعات مکانی نهفته در اسناد مکانی و پیاده‌سازی و ارزیابی نگرش یکپارچه در بازیابی این اطلاعات می‌باشد. نگرش کلی در بازیابی اطلاعات مکانی به نحوی است که این اطلاعات از طریق ارتباطی که به اطلاعات غیر مکانی دارند استخراج می‌شوند، در حالی که در اسناد مکانی موجود اطلاعات مکانی و غیرمکانی به صورت یکپارچه ذخیره می‌گردند. در پژوهش‌های پیشین اسناد مکانی و اطلاعات موجود در آن‌ها کمتر مورد توجه قرار گرفته است. منظور از نگرش یکپارچه در بازیابی اطلاعات مکانی، استخراج اطلاعات مکانی و توصیفی موجود در اسناد مکانی به صورت یکپارچه و همزمان می‌باشد. اجزای تشکیل دهنده سیستم مبتنی بر پژوهش حاضر شامل خزنده، پایگاه داده و واسط کاربری می‌باشد. در بخش خزنده، اسناد مکانی کشف شده و متن این اسناد برای استخراج اطلاعات تجزیه می‌شود. پایگاه داده در این سیستم وظیفه ذخیره و شاخص‌گذاری اطلاعات استخراج شده توسط خزنده را برعهده دارد و در نهایت واسط کاربری تعامل بین سیستم و کاربر را فراهم می‌کند. این سیستم به صورت آزمایشی برروی یک کارساز کاربری به عنوان یک شبیه سازی از فضای وب پیاده‌سازی شده است. پژوهش پیش رو با پیاده‌سازی نگرش یکپارچه، اطلاعات مکانی را از اسناد مکانی بازیابی می‌کند و به این ترتیب گام مؤثری در بهبود کارایی موتورهای جستجوی مکانی برمی‌دارد.

واژگان کلیدی: وب مکانی، موتورهای جستجوی مکانی، اسناد مکانی، خزنده، جی ام ال GML

* نویسنده رابط

۱- مقدمه

مفهوم دسترسی آزاد به اطلاعات ایجاب می‌کند که تمام کسانی که به نحوی می‌توانند از اطلاعاتی استفاده کنند به آن دسترسی داشته باشند، به گونه‌ای که با داشتن کمترین دانش فنی در مورد مجموعه داده‌ها، به صورت انتخابی امکان بازیابی اطلاعات برای آنها فراهم باشد. به صورت کلی امکان دسترسی انتخابی به اطلاعات بر روی وب از طریق موتورهای جستجو فراهم می‌شود، اما در مواردی که نیاز کاربر در برگیرنده اطلاعات مکانی نیز باشد وظیفه جستجو پیچیده‌تر می‌شود و احتیاج به توانایی‌های خاصی در بخش جستجوگر است.

امروزه اکثر صفحات وب شامل اطلاعاتی هستند که به نحوی با موقعیت مکانی پیوند خورده‌اند. این اطلاعات موقعیت معمولاً توسط موتورهای جستجوی سنتی نادیده گرفته می‌شوند. جستارهایی که بصورت ترکیبی از موقعیت و متن می‌باشند به نام جستارهای مکانی-متنی شناخته می‌شوند . با ورود اطلاعات مکانی به دنیای وب، موتورهای جستجو در یک روند مشخص از پردازش متنی و چند رسانه‌ای به سمت پردازش اطلاعات مکانی و جستارهای مکانی-متنی گرایش پیدا کرده‌اند. تمام جنبه‌های فعالیت بشر ریشه در فضای مکانی دارد. به عنوان یک نتیجه، بسیاری از اسناد شامل ارجاعاتی به متون مکانی می‌باشند که معمولا این ارجاعات به وسیله نام مکان‌ها صورت می‌پذیرد. این رخداد رایج در اسنادی که در دنیای وسیع وب ذخیره و بازیابی شده‌اند، نمود پیدا می‌کند. اگر یک کاربر موتور جستجو بخواهد منابعی را پیدا کند که موضوع اصلی به یک مکان مشخص مرتبط است، می‌تواند مکان موردنظر خود را در پرسش[۱] موتور جستجو لحاظ کند. رفتار موتورهای جستجوی مرسوم در مواجه با نام مکان‌ها به همان صورت سایر کلمات کلیدی می‌باشد و اسنادی را که شامل نام مشخص شده می‌باشند، بازیابی می‌کنند. اگرچه برای بعضی اهداف این امر کافی بنظر می‌رسد، اما موقعیت‌های بسیاری وجود دارد که کاربر علاقه دارد اسناد مربوط به آن ناحیه که به وسیله نام مکان مشخص شده است را ببیند، درحالیکه این اسناد شامل آن نام نمی‌باشند. در اینجا نیاز به یک موتور جستجوی مکان آگاه[۲] می‌باشد که بتواند ذکر یک مکان در یک پرسش را با هوشمندی تفسیر کند، به گونه‌ای که نتایج با کیفیت بالا بازیابی شوند(۱). مطالعات مختلف نشان می‌دهد که حداقل یک پرسش از پنج پرسش در موتورهای جستجوی عمومی یک بعد مکانی دارد و می‌تواند به عنوان یک پرسش مکان‌مبنا در نظر گرفته شود(۲, ۳).

امروزه دنیای وب شامل اطلاعات گسترده‌ای می‌باشد که برخی از این اطلاعات می‌توانند زمین مرجع باشند. آکادمی ملی علوم آمریکا تخمین زده‌است که ۸۰ درصد اطلاعات روی وب یک جز مکانی دارند(۴). این جز مکانی می‌تواند شامل اطلاعات مختصات مانند طول و عرض جغرافیایی[۳] و انواع مختلف سیستم تصویرهایشان، آدرس‌های پستی که می‌توانند زمین‌کدیابی[۴] شوند و اطلاعات وابسته به مسافت و جهت باشد. همچنین یک مطالعه در مطبوعات گزارش می‌کند که از بین ۲۵۰۰ جستار، ۱۸,۶ درصد آنها شامل یک گزاره جغرافیایی و ۱۴,۸ درصد آنها شامل نام یک مکان می‌باشند(۵). بنابراین چگونگی استخراج موقعیت‌ها از صفحات وب و سپس استفاده از آن‌ها در فرآیند جستجوی وب تبدیل به یک بحث حیاتی در پژوهش‌های اخیر در موتورهای جستجو شده است.

روزانه داده‌های زیادی توسط ماهواره‌ها، حسگرهای زمینی، شبیه‌سازی‌های کامپیوتری و دستگاه‌های موبایل تولید می‌شود(۶). قسمت قابل توجهی از این داده‌ها به مکان مرتبط می‌شوند. بدیهی‌ست با وجود چنین ابزارهایی، امروزه ایجاد داده مکانی به مراتب آسان‌تر از گذشته شده است. در این راستا کنسرسیوم مکانی باز[۵] (OGC) بطور فزاینده‌ای پروتکل‌هایی[۶] را برای استاندارد کردن چگونگی ذخیره و اشتراک‌گذاری داده‌های مکانی بر روی وب ارائه می‌کند(۷).

پیشرفت کمی اطلاعات با پیچیدگی‌هایی در فرآیند ذخیره سازی و بازیابی اطلاعات همراه خواهد بود. با اضافه شدن گسترده داده مکانی به داده‌های موجود در فضای وب روند بازیابی این نوع داده‌ای خاص با چالش‌های بسیاری مواجه شده است. یکی از اولین و مهمترین

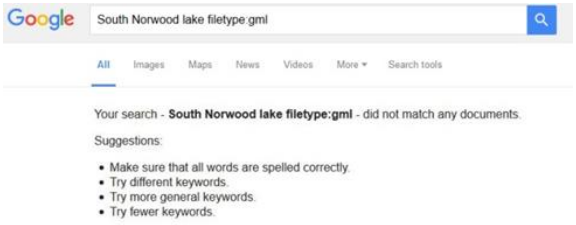
---

۱ query

۲ Spatially-aware
۳ Longitud and latitud
٤ geocoding
٥ Open Geospatial Consortium
٦ protocol

چالش‌های پیش رو در این زمینه نحوه ذخیره، اشتراک-گذاری و انتقال داده‌های مکانی می‌باشد. پس از فراگیر شدن داده‌های مکانی، سازمان‌ها، انجمن‌ها، کاربران اینترنتی و بسیاری از ارگان‌های مرتبط با این شاخه از علم شروع به تهیه و تنظیم استانداردها، ساختارها و قالب‌های متفاوتی برای ذخیره و به اشتراک‌گذاری این نوع از داده‌ها کردند. اسنادی که برای ذخیره داده از اینگونه ساختارها و استانداردها تبعیت می‌کنند به اسناد مکانی موسوم هستند. ویژگی قابل توجه اسناد مکانی قابلیت نگهداری اطلاعات توصیفی و غیر مکانی در کنار اطلاعات مکانی می‌باشد. این اطلاعات غیرمکانی به نوعی اطلاعات مکانی موجود را توصیف کرده و درک بهتری از این اطلاعات را به استفاده کنندگان ارائه می‌دهند. پردازش متن اینگونه داده‌ها و استخراج اطلاعات مکانی موجود در آن‌ها می‌تواند در بازیابی و ارائه نتایج در یک موتور جستجوی مکانی بسیار موثر باشد. تاکنون کارهای زیادی برروی پردازش متن اسناد زبان نشانه‌گذاری ابرمتنی[۱] (HTML) صورت گرفته است، اما خلاء پردازش متون اسناد مکانی مثل زبان نشانه‌گذاری جغرافیایی[۲] (GML) در موتورهای جستجو احساس می‌شود. تمرکز پژوهش ارائه شده در این مقاله بر روی بازیابی اطلاعات از اسناد مکانی می‌باشد تا با ایجاد یک نگرش یکپارچه اطلاعات مکانی و غیر مکانی مربوط به عوارض از اسناد مربوطه به درستی استخراج شوند. نگرش کلی در بازیابی اطلاعات مکانی به نحوی است که این اطلاعات از طریق اتصالی که به اطلاعات غیرمکانی دارند استخراج می‌شوند، در حالی که در اسناد مکانی موجود اطلاعات مکانی و غیر مکانی به صورت یکپارچه ذخیره می‌گردند.

هنگامی که کاربری به دنبال یک سند مکانی باشد، از آنجاییکه موتورهای جستجوی عمومی هیچ حساسیتی نسبت به محتوای موجود در اسناد GML ندارند، کاربر قادر به یافتن اینگونه اسناد از طریق موتورهای جستجو نمی‌باشد. شکل۱ پرسشی را در موتور جستجوی Google نمایش می‌دهد که طی آن کاربر قصد دارد اسناد GML را که شامل اطلاعات مرتبط با پرسش هستند بیابد.



شکل۱- نمونه‌ای از پاسخ Google به یک پرسش مکانی

برای جستجوی یک قالب خاص در Google می‌توان از عبارت "filetype:" قبل از مخفف آن قالب استفاده کرد. به عنوان مثال عبارت "filetype:pdf" بعد از یک پرسش در Google بیانگر این است که کاربر به دنبال اسناد مرتبط با پرسش با قالب pdf می‌گردد. بنابراین عبارت "filetype:gml" به این معنی است که کاربر به دنبال اسناد با قالب GML تحت موضوع مورد پرسش می‌باشد. همانطور که در شکل مشخص است موتور جستجوی Google در قبال پرسش " South Norwood Lake filetype:gml" پیغامی را مبنی بر عدم تطبیق هیچ سندی با پرسش کاربر در خروجی نمایش می‌دهد. نکته مهمی که از این اتفاق برداشت می‌شود عدم توجه Google به اسناد با قالب GML می‌باشد. لازم به ذکر است که سندی با این مشخصات در فضای وب موجود می‌باشد. شکل۲ سندی را نمایش می‌دهد که عبارت جستجو شده عیناً در آن ذکر شده است.

شکل۲- نمونه‌ای از یک سند GML موجود در وب

سیستم پیشنهادی از سه جز اصلی خزنده[۳]، پایگاه داده و واسط کاربری تشکیل شده است. خزنده در این سیستم وظیفه یافتن و تجزیه کردن اسناد مکانی را بر عهده دارد و به نوعی نوآوری اصلی پژوهش محسوب می‌شود. پایگاه داده برای ذخیره اطلاعات استخراج شده از

۱ HyperText Markup Language
۲ Geography Markup Language
۳ Crawler

اسناد مکانی توسط خزنده مورد استفاده قرار می‌گیرد و واسط کاربری تعامل بین کاربر و سیستم را فراهم می‌کند و کاربر از طریق آن پرسش خود را به سیستم ارسال کرده و نتایج حاصل از جستجوی خود را مشاهده می‌کند. در بخش ۳ اجزای سیستم پیشنهادی به تفصیل مورد بررسی قرار خواهند گرفت.

پیاده سازی این سیستم در زبان برنامه نویسی جاوا انجام شده است. برای خزنده در این سیستم از یک خزنده متن باز به نام کراولر۴جی[۱] که در جاوا نوشته شده است و از پایگاه داده مای‌اس‌کیوال[۲] در بخش پایگاه داده این سیستم استفاده شده است. پیاده‌سازی در دو مرحله برون خط و برخط به صورت جداگانه در بخش ۴ مورد بررسی قرار خواهد گرفت.

در ادامه ابتدا به پژوهش‌هایی که در حوزه موتورهای جستجوی مکانی انجام گرفته است به اختصار اشاره خواهد شد. سپس به بررسی روش انجام پژوهش و مبانی تئوری آن پرداخته می‌شود. در بخش بعدی با تبیین پیاده‌سازی سیستم پیشنهادی، نتایج حاصل از پژوهش ارائه خواهد شد و در نهایت در بخش آخر ضمن جمع‌بندی مطالب ارائه شده پیشنهاداتی جهت تکمیل پژوهش حاضر در آینده مورد بررسی قرار خواهد گرفت.

۲- پیشینه پژوهش

برخلاف جستجوهای عمومی، در جستجوهای مکان-مبنا انتظار می‌رود اسنادی بازیابی و رتبه‌بندی شوند که نه فقط به موضوع پرسش ، بلکه بصورت جغرافیایی به مکانی که با پرسش در ارتباط است مرتبط باشند. مسائل زیادی بر سر راه توسعه موتورهای جستجوی مکانی موثر وجود دارد و تاکنون هیچ موتور جستجوی مکان‌مبنای جهانی گزارش نشده است(۸). شرکت‌های موتور جستجو شروع به توسعه و ارائه خدمات مکان‌مبنا کرده‌اند، با این وجود همچنان از نظر جغرافیایی دارای محدودیت‌هایی می‌باشند و به اندازه موتورهای جستجوی عمومی موفق و مشهور نشده‌اند.

دو نمونه از موتورهای مکان‌آگاه تجاری گوگل مپ[۳] و یاهو لوکال[۴] هستند. این موتورها از دو واسط کاربری کلیدواژه‌مبنا و نقشه‌مبنا برای یافتن اطلاعات وابسته به یک آدرس یا موقعیت مشخص استفاده می‌کنند. به صورت اسمی این موتورهای جستجو کل جهان را پوشش می‌دهند، با این وجود از آنجا که فرهنگ‌های جغرافیایی و نقشه‌های رقومی برای تمام کشورها در دسترس نیستند، پوشش این موتورها اغلب به چند کشور خاص محدود می‌باشد(۲).

جونز و همکاران در سال ۲۰۰۴ به معرفی یک موتور جستجوی مکان‌آگاه به نام اسپیرایت[۵] پرداخته‌اند(۱). موتور جستجوی اسپیرایت یک بستر آزمایشی برای توسعه فناوری جستجوی وب فراهم کرده است که بصورت خاص برای دسترسی به اطلاعات جغرافیایی طراحی شده است. اجزای اصلی این سیستم شامل واسط کاربری، هستی‌شناسی جغرافیایی، توابع نگهداری و بازیابی برای یک مجموعه آزمایشی از اسناد وب، شاخص‌های متنی و مکانی، موتور جستجوی مرکزی، رتبه‌دهی و استخراج فراداده می‌باشد. واسط کاربری به کاربر اجازه می‌دهد یک موضوع مورد علاقه و یک موقعیت جغرافیایی را تعیین کند. اولین قیود برای تعیین یک پرسش، یک واسط متنی ساختارمند، یک واسط متنی آزاد و یک نقشه می‌باشد. واسط متنی ساختارمند به کاربر اجازه می‌دهد که موضوع پرسش، یک نام مکان و یک رابطه مکانی برای نام مکان را تعیین کند.

یک راه حل برای جستجوی داده‌های مکانی برروی وب استفاده از روش خزنده وب می‌باشد که می‌تواند تمام وب را برای اطلاعات مکانی جستجو کند(۹). خزنده‌های وب نرم افزارهای اینترنتی هستند که وب را با هدف شاخص‌گذاری محتویات مرور می‌کنند. خزنده‌ها در ابتدا به لیستی از آدرس‌های وب[۶] مشخص سر می‌زنند و سپس این لیست را با شناسایی فراپیوندها[۷] و اضافه کردن آدرس‌های وب آن‌ها به آن گسترش می‌دهند. در ادامه به سایت‌های داخل این لیست به صورت بازگشتی براساس یک مجموعه تعریف شده از قوانین برای تعیین اینکه آیا هنوز شامل معیارهای جستجو هستند سر زده می‌شود. با این کار، خزنده یک لیست جدید از سایت‌هایی که تعدادی

۱ Crawler4j
۲ MySQL
۳ Google Maps
۴ Yahoo Local
۵ SPIRIT
۶ URL
۷ hyperlinks

معیار دارند فراهم می‌کند، در حالی که آدرس‌های وبی که در طول زمان غیر مرتبط و یا معیوب شده‌اند حذف می‌گردند. بن و همکاران در سال ۲۰۱۴ یک موتور جستجوی مکانی (GSE)[۱] ارائه کرده‌اند که از یک خزنده وب برمبنای موتور جستجوی گوگل به منظور جستجو در وب برای داده‌های مکانی استفاده می‌کند(۶). مکانیزم خوراک‌دهی اولیه خزنده در GSE واژه‌های جستجویی که توسط کاربر داده شده‌اند را با کلمات کلیدی از پیش تعیین شده که سرویس‌های داده‌های مکانی را شناسایی می‌کنند ترکیب می‌کند. این سیستم از خدمات نقشه وبی (WMS)[۲]، خدمات آرک جی آی اس[۳] و وبسایت‌هایی که داده مکانی دارند، پشتیبانی می‌کند. در واقع GSE یک برنامه وبی می‌باشد که یک پایگاه داده قابل جستجو از کارسازهای[۴] مکانی و لایه‌ها را با یک خزنده وب برای تعیین داده‌های مکانی جدید و نیز به‌روزرسانی پایگاه داده مکانی موجود ترکیب می‌کند. در GSE از یک خزنده وب یک سطحی[۵]که برروی موتور جستجوی گوگل ساخته شده است استفاده می‌شود. واژه یک سطحی به این موضوع اشاره دارد که خزنده فقط خدمات بازگردانده شده در طول جستجو را پیمایش می‌کند و نسبت به پیوندهایی که در داخل این سرویس‌ها وجود دارند بی توجه است. علی‌رغم اینکه GSE یک موتور جستجوی مکانی خزنده‌مبنا می‌باشد، ولی همچنان به متن اسناد مکانی توجه چندانی ندارد و صرفاً به دنبال خدمات مکانی بر روی وی می‌گردد. تک سطحی بودن خزنده GSE به خوبی مبین این موضوع است که هدف آن یافتن اسناد مکانی و تجزیه متن آن‌ها نمی‌باشد.

امروزه استخراج اطلاعات موقعیت از صفحات وب به یک موضوع چالشی در زمینه جستجوگرهای وب تبدیل شده است. یک رویکرد در این زمینه توسط ژائو و همکاران در سال ۲۰۱۴ ارائه شده است(۱۰). نویسندگان یک چهارچوب برای استفاده از اطلاعات موقعیت جهت جستجوی وب ارائه می‌دهند. چهارچوب پیشنهادی شامل یک مرحله برون خط برای استخراج موقعیت متمرکزشده از صفحات وب پویش شده و نیز یک مرحله رتبه‌دهی برخط جهت تشکیل رتبه‌دهی مکان‌آگاه برای نتایج جستجو می‌باشد. موقعیت متمرکزشده یک صفحه وب به موقعیت‌های مناسب که با صفحه وب آمیخته شده‌اند اشاره دارد. در مرحله برون خط موقعیت‌های متمرکزشده و کلیدواژه‌ها از صفحات وب کشف شده و هرکلیدواژه با موقعیت‌های متمرکزشده تشکیل یک مجموعه از زوج مرتب‌های <کلیدواژه،موقعیت> را می‌دهند. در مرحله پردازش پرسش برخط، کلیدواژه‌ها از پرسش‌ها استخراج می‌شوند و رتبه کسب شده براساس ارتباط موقعیت و قیود موقعیت برای هر کلیدواژه پرسش محاسبه می‌شود.

آمیتای و همکاران در سال ۲۰۰۴ یک الگوریتم ابتکاری چهارگامی به نام وب-ا-ور[۶] برای تعیین موقعیت‌های متمرکزشده برای صفحات وب ارائه کرده‌اند که در آن تمام اسم‌ها به یک موقعیت با یک درجه اطمینان اختصاص داده می‌شوند(۱۱). براساس آن درجه اطمینان و نیز سایر پارامترها نظیر تکرار و ارتباط مکانی، موقعیت متمرکزشده در یک صفحه وب استخراج می‌شود. الگوریتم وب-ا-ور اشارات به نام مکان‌ها را موقعیت‌یابی و ارتباط هر نام با مکان را مشخص می‌کند. همچنین این سیستم به هر صفحه یک تمرکز جغرافیایی اختصاص می‌دهد. تمرکز جغرافیایی یک مکان است که یک صفحه تماماً در مورد آن بحث می‌کند. از نظر نویسندگان اکثر اسامی مکان‌ها و اکثریت قابل توجهی از اسامی یافت‌شده در وب دارای ابهام هستند. آن‌ها دو نوع ابهام برای مکان‌های جغرافیایی متصور شده‌اند. ابهام زمین/غیر-زمین[۷] به حالتی اطلاق می‌شود که یک نام مکان، معنی یا معانی غیر جغرافیایی دیگری دارد(مثل واشنگتن به عنوان نام یک شخص و همچنین یک شهر). ابهام دیگری که نویسندگان از آن به عنوان زمین/زمین[۸] یاد کرده‌اند به حالتی گفته می‌شود که دو منطقه جدا از هم یک نام مشترک دارند(به عنوان مثال ایالات متحده آمریکا ۲۴ شهر به نام پاریس دارد).

در تمامی پژوهش‌هایی که تاکنون برروی موتورهای جستجوی مکان‌آگاه صورت گرفته است تمرکز بر روی نمایش نقشه و یا کشف موقعیت از صفحات وب بوده است و بازیابی اطلاعات از متن اسناد مکانی مثل GML با ایجاد

۱ Geospatial Search Engine
۲ Web Map Service
۳ ArcGIS Services
٤ server
٥ single –level

۶ Web-a-Where
۷ Geo/non-Geo
۸ Geo/Geo

یک نگرش یکپارچه در مورد داده‌های مکانی و غیرمکانی کمتر مورد توجه قرار گرفته است. پژوهش ارائه شده در این مقاله با پیاده‌سازی نگرش یکپارچه، اطلاعات مکانی را از اسناد مکانی بازیابی می‌کند و گام موثری در دسترسی انتخابی به داده‌های مکانی موجود در وب برمی‌دارد.

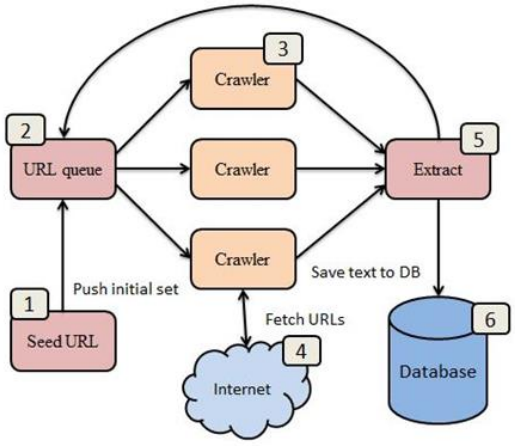
۳- روش پژوهش

در این بخش روش پیشنهادی برای ایجاد سیستم جستجوی مکان‌مبنا ارائه خواهد شد. ابتدا با بررسی روش‌های جستجو در اسناد وب به تبیین روش جستجوی خزنده‌مبنا پرداخته خواهد شد. سپس نگرش یکپارچه در بازیابی اطلاعات مکانی ارائه شده و با نگرش کلی مقایسه می‌شود. در ادامه نیز با بررسی اجمالی سیستم حاضر اجزای مختلف آن تبیین خواهند شد.

۳-۱- جستجوی خزنده مبنا

تاکنون روش‌های مختلفی برای جستجوی اسناد موجود در سطح وب ارائه شده است. کامل‌ترین دسته‌بندی که برای موتورهای جستجو وجود دارد آنها را به چهار دسته موتورهای جستجوی پیمایشی، فهرست تکمیل دستی، موتورهای جستجوی ترکیبی با نتایج مختلط و فراجستجوگرها تقسیم‌بندی می‌کند. موتورهای جستجوی پیمایشی وب را توسط تعدادی خزنده پیمایش کرده و اطلاعات را ذخیره می‌کنند. در حالی که فهرست‌های تکمیل‌دستی وابسته به کاربرانی است که آن را تکمیل می‌کنند. صاحبان وبسایت‌ها صفحه مورد نظر خود را به همراه توضیحی کوتاه در فهرست ثبت می‌کنند یا این کار توسط ویراستارهایی که برای آن فهرست در نظر گرفته شده، انجام می‌شود. موتورهای جستجوی ترکیبی با نتایج مختلط هر دو حالت پیمایشی و تکمیل دستی را در کنار هم نمایش می‌دهند. فراجستجوگرها نیز که قدمت چندانی ندارند از ترکیب کردن نتایج حاصل از چندین موتورجستجو استفاده می‌کنند.

همانطور که اشاره شد یکی از روش‌های جستجو در میان مجموعه‌ای از اسناد وب استفاده از خزنده می‌باشد. شکل۳ طرحواره عملکرد یک خزنده را نمایش می‌دهد.



شکل۳- طرحواره عملکرد یک خزنده

همانطور که در شکل مشخص است خزنده پس از دریافت خوراک اولیه وارد صفحات اولیه می‌شود و متن این صفحات را استخراج کرده و در پایگاه داده ذخیره می‌کند. همچنین به تمامی پیوندهای موجود در این صفحات سر می‌زند و این فرآیند را تا جایی ادامه می‌دهد که هیچ پیوند جدیدی پیدا نشود.

در این روش خزنده پس از یافتن اسناد آن‌ها را به تجزیه‌کننده تحویل می‌دهد. تجزیه‌کننده متن هر سند را می‌خواند و سپس آن را براساس کلمات کلیدی دسته بندی کرده و در مخزن ذخیره می‌کند. همچنین تجزیه‌کننده خطاهای موجود در متن اعم از خطاهای دستوری و نگارشی را تشخیص می‌دهد و اقدام به تصحیح آن‌ها می‌کند.

در گام بعدی، کلیدواژه‌های تجزیه شده از متن اسناد در پایگاه داده شاخص‌گذاری شده و بسته به ماهیت‌شان در جداول پایگاه داده ذخیره می‌شوند. حال سیستم مجموعه‌ای از اسناد را شامل می‌شود که هر سند به یکسری از کلیدواژه‌ها پیوند خورده است. تمام مراحلی که تا اینجا اشاره شد به صورت برون خط می‌باشد، به این معنی که عملیات مذکور بدون در نظر گرفتن هیچ پرسشی از سوی کاربر انجام می‌شود و سیستم موظف است در بازه‌های زمانی مشخص این عملیات را تکرار کرده و اطلاعات موجود را به روز رسانی کند تا بتواند تغییرات به وجود آمده در هر سند را در پایگاه داده اعمال کند، ضمن اینکه اسناد جدید را نیز مورد بررسی قرار دهد. در واقع مراحل فوق مستقل از درخواست‌های کاربران خوانده می‌شود.

پس از پایان بخش مستقل از کاربر، گام‌های وابسته به کاربر آغاز می‌شوند. گام نخست در این بخش پردازش پرسش کاربر می‌باشد. کاربر پرسش خود را از طریق واسط کاربری به سیستم ارسال می‌کند. سپس سیستم با پردازش این پرسش اقدام به ارائه نتیجه به کاربر می‌کند. همانطور که می‌دانیم اولین شرط برای ارائه یک جواب قابل قبول به یک سوال، فهم درست آن سوال است. موتورهای جستجو نیز از این قاعده مستثنی نیستند و شرط لازم برای ارائه نتایج قابل قبول به یک پرسش در موتورهای جستجو فهم پرسش توسط سیستم می‌باشد. پس از آنکه سیستم پرسش را پردازش کرد در پایگاه داده خود به دنبال اسنادی می‌گردد که با درخواست کاربر همخوانی دارند.

حال فرض کنیم سیستم در مخزن خود چندین سند همخوان با نیاز کاربر پیدا می‌کند. مسئله مهم در اینجا این است که کدام سند همخوانی بیشتری با درخواست کاربر دارد یا به عبارتی کدام جواب صحیح‌تر است. به این مسئله باید در گام بعدی پاسخ داده شود، جایی که با استفاده از الگوریتم‌های رتبه‌دهی هر سند همخوان حائز یک رتبه می‌شود و براساس این رتبه‌ها ترتیب اسناد خروجی که به کاربر ارائه می‌شوند مشخص می‌گردد. به این ترتیب پاسخ-های صحیح‌تر از نظر سیستم (و نه لزوماً از نظر کاربر) در رتبه‌های بالاتر فهرست قرار می‌گیرند. اینکه نظر سیستم تا چه اندازه با نظر کاربر تطابق داشته باشد به دقت الگوریتم رتبه‌دهی مورد استفاده مربوط می‌باشد.

۳-۲- نگرش یکپارچه در بازیابی اطلاعات مکانی

یک سند حاوی اطلاعات مکانی بر روی وب می‌تواند شامل اطلاعات مکانی زیادی از جمله عوارض، مختصات، روابط مکانی و توصیفات مکانی باشد. پردازش متن اینگونه داده‌ها و استخراج اطلاعات مکانی موجود در آن‌ها می‌تواند در بازیابی و ارائه نتایج در یک موتور جستجوی مکانی بسیار موثر باشد. تاکنون کارهای زیادی برروی پردازش متن اسناد HTML صورت گرفته است، اما خلاء پردازش متون اسناد مکانی مثل GML در موتورهای جستجو احساس می‌شود. نگرش کلی در بازیابی اطلاعات مکانی به نحوی است که این اطلاعات از طریق اتصالی که به اطلاعات غیرمکانی دارند استخراج می‌شوند، در حالی که در اسناد مکانی موجود اطلاعات مکانی و غیر مکانی به صورت یکپارچه ذخیره می‌گردند. نگرش یکپارچه در قبال اطلاعات مکانی و غیر مکانی را می‌توان بهترین راه برای بازیابی اطلاعات مکانی دانست و برای ایجاد این نگرش راهی به جز استفاده از اسناد مکانی وجود ندارد.

شکل ۴ متن یک سند GML را نمایش می‌دهد. همانطور که مشاهده می‌شود اطلاعات مکانی (بیضی خط چین) و غیرمکانی (بیضی ممتد) در اینگونه اسناد در کنار هم نگهداری می‌شوند.

شکل۴- متن یک سند GML: بیضی‌های خط چین و ممتد به ترتیب بیانگر اطلاعات مکانی و غیر مکانی می‌باشند

در مقابل شکل ۵ متن یک سند HTML را نمایش می‌دهد. همانطور که در شکل مشخص است واژه "تهران" در این سند به تنهایی مبین یک داده مکانی نیست و با ارتباطی که با داده‌های غیر مکانی موجود در متن دارد می‌توان آن را یک داده مکانی در نظر گرفت.

شکل۵- یک سند HTML: عناصر مکانی در این اسناد با اتصال به عناصر غیرمکانی معنی پیدا می‌کنند

با توضیحات داده شده ارزش اسناد مکانی و اطلاعات نهفته در آن‌ها بیش از پیش نمایان می‌شود. همانطور که پیش‌تر ذکر شد، بسیاری از متخصصین در کاربردهای مختلف به داده مکانی خام نیاز دارند. در حالی که موتورهای

جستجوی مکانی موجود قادر به بازیابی این داده‌ها برای این کاربردها نیستند و صرفاً به نمایش نقشه بسنده می‌کنند و طبیعتاً با توجه به گستردگی دنیای وب، بازیابی این داده‌ها از طریق موتوهای جستجوی عمومی نیز برای این کاربران کار ساده‌ای به نظر نمی‌رسد. لذا ایجاد بستری برای یافتن داده مکانی خام گام موثری در پیشبرد فرآیند دسترسی انتخابی به اطلاعات در دنیای وب خواهد برداشت.

سیستم حاضر بر مبنای اطلاعات مکانی ایجاد شده است و دسترسی به این اطلاعات برای کاربران با استفاده از این سیستم بسیار آسان است. برای کاربری که به اطلاعات مختصاتی از شهر تهران نیاز دارد، یک راه حل این است که در موتورهای جستجوی مکانی مثل گوگل مپ به دنبال شهر تهران بگردد و پس از یافتن آن مختصات نقاط مورد نظر را یادداشت کند. این راه حل عملی می‌باشد، اما راه حل ساده‌تر این است که بتواند به راحتی به اسناد GML که به تهران مربوط می‌باشند دسترسی داشته باشند. ضمن اینکه علاوه بر اسناد GML بتواند اطلاعات توصیفی مختلفی درباره تهران دریافت کند و حتی قادر به مشاهده یک گستره از تهران باشد. از آنجاییکه در راه حل دوم علاوه بر مشاهده نقشه اطلاعات کامل‌تری درباره منطقه مورد جستجو در اختیار کاربر قرار می‌گیرد، این راه حل مورد ترجیح است.

۳-۳- بررسی اجمالی سیستم

پیکره اصلی سیستم پیشنهادی متشکل از سه جزء زیر می‌باشد:

۱- خزنده: نوآوری اصلی این سیستم در خزنده گنجانده شده است و از این رو می‌توان خزنده را مهمترین جزء سیستم پیشنهادی دانست. خزنده در این سیستم به جای یافتن اسناد سنتی مثل HTML و یا قالب‌های متنی دیگر از قبیل پی دی اف به دنبال اسناد مکانی می‌گردد. علاوه بر این خزنده اقدام به تجزیه متن اسناد مکانی کرده و اطلاعات مکانی و غیر مکانی موجود در آنها را استخراج می‌کند. سپس این اطلاعات را به تفکیک در جداول پایگاه داده ذخیره می‌کند. در این مرحله از پژوهش، از بین تمامی اسناد مکانی، قالب GML به دلیل فراگیر بودن انتخاب شده است، اما با اندکی تغییر در بخش خزنده نتایج به دست آمده به راحتی قابل اعمال به سایر قالب‌های مکانی بر روی وب هستند. بنابراین می‌توان دو وظیفه زیر را به عنوان وظایف اصلی خزنده در این سیستم نام برد:

- تشخیص اسناد GML از بین اسناد با قالب‌های متفاوت
- تجزیه متن اسناد GML و استخراج اطلاعات خام مکانی به همراه اطلاعات غیر مکانی موجود در این اسناد

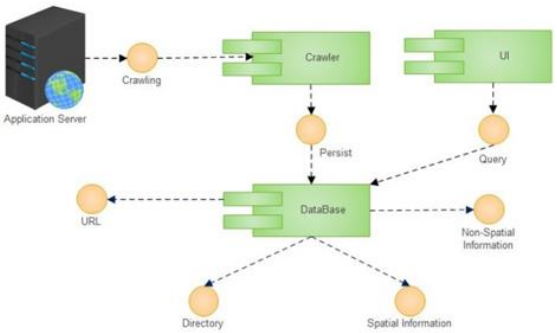
۲- پایگاه داده: وظیفه اصلی پایگاه داده ذخیره و نگهداری اطلاعات جمع آوری شده توسط خزنده می‌باشد. ضمن اینکه هر بار خزنده عملیات جستجو رو تکرار می‌کند پایگاه داده موظف است خود را به‌روز کند و تغییرات به وجود آمده در اسناد موجود را اعمال کند. همچنین در صورت یافتن اسناد جدید پایگاه داده موظف است اطلاعات این اسناد را به جداول خود اضافه کند.

۳- واسط کاربری: واسط کاربری تعامل بین کاربر و سیستم را فراهم می‌کند. کاربر پرسش خود را در واسط کاربری وارد کرده و آن را به سیستم می‌فرستد و سیستم پس از طی فرآیندهای جستجو، نتایج را به ترتیب میزان همخوانی با نیاز کاربر که توسط الگوریتم رتبه‌دهی مشخص می‌شود به واسط کاربری برمی‌گرداند تا در اختیار کاربر قرار گیرد.

شکل۶ معماری سیستم پیشنهادی را به صورت خلاصه نمایش می‌دهد. در پیاده‌سازی نسخه آزمایشی سیستم پیشنهادی از یک کارساز کاربردی[۱] به عنوان یک شبیه‌سازی از دنیای وب استفاده شده است و اسناد در آن نگهداری می‌شوند. همانطور که در شکل مشخص است خزنده اسناد GML را یافته و پس از تجزیه این اسناد، آنها را به تفکیک محتوا در جداول مختلف پایگاه داده ذخیره می‌کند. پایگاه داده در این سیستم متشکل از ۳ جدول پایه و همچنین یک جدول اضافی می‌باشد که برای نسخه آزمایشی طراحی شده است و در نسخه نهایی حذف خواهد شد. جداول پایه عبارتند از جدول اطلاعات مکانی، جدول اطلاعات غیر مکانی و جدول مربوط به نشانی وب و جدول اضافی آدرس اسناد بر روی حافظه را نگهداری می‌کند. واسط کاربری نیز واسطه بین کاربر و سیستم است و پرسش کاربر از طریق آن به سیستم ارسال می‌شود و در نهایت نتایج نیز از طریق آن به کاربر نمایش داده می‌شوند.

---

۱ Application server



شکل۶- معماری سیستم پیشنهادی

جستجو در این سیستم به دو روش کلیدواژه مبنا و مختصات مبنا امکان‌پذیر می‌باشد. در جستجو بر مبنای کلیدواژه کاربر کلیدواژه مورد نظر خود را در واسط کاربری وارد می‌کند. واسط کاربری این کلیدواژه را برای انجام عملیات بعدی به سیستم ارسال می‌کند. در جستجوی مختصات مبنا کاربر به جای وارد کردن کلیدواژه، مختصات نقطه‌ای را وارد می‌کند که می‌خواهد در مورد عوارض اطراف آن نقطه اطلاعاتی کسب کند. پس از آنکه واسط کاربری مختصات نقطه مذکور را به سیستم ارسال کرد، سیستم شروع به محاسبه فاصله این نقطه با عوارض موجود در اسناد مکانی می‌کند و هر عارضه‌ای که فاصله آن تا نقطه مورد نظر کاربر از یک حد آستانه کمتر بود، سند حاوی آن عارضه به کاربر برمی‌گردد. حد آستانه‌ای که در این روش مورد استفاده قرار می‌گیرد توسط کاربر در واسط کاربری وارد می‌شود. در واقع کاربر مشخص می‌کند علاقه‌مند به دریافت اسناد تا چه فاصله‌ای از نقطه مورد نظرش می‌باشد. به این ترتیب تمامی اسنادی که حداقل یکی از عوارض موجود در آن داخل دایره‌ای به شعاع حد آستانه مشخص شده توسط کاربر باشد به کاربر برمی‌گردد. در مورد عوارض چندضلعی از مستطیل محاط به جای فاصله استفاده می‌کنیم. به این صورت که چندضلعی‌هایی که نقطه مورد نظر کاربر داخل مستطیل محاط آن‌ها باشد به کاربر برمی‌گردد.

۳-۴- بررسی مراحل اجرایی سیستم

همانطور که پیش‌تر اشاره شد هر موتور جستجو بر مبنای خزنده از دو مرحله برون خط و برخط تشکیل شده است. سیستم پیشنهادی ما نیز از این قاعده مستثنی نیست. در این سیستم مرحله برون خط عبارت است از فرآیند جستجوی خزنده و یافتن و تجزیه کردن اسناد GML و ذخیره اطلاعات استخراج شده در پایگاه داده و همچنین بروزرسانی پایگاه داده در بازه‌های زمانی مختلف. مرحله برخط نیز به ارسال پرسش توسط کاربر به سیستم از طریق واسط کاربری و محاسبه رتبه اسناد همخوان با پرسش مذکور توسط الگوریتم رتبه‌دهی خلاصه می‌شود. در ادامه به بررسی دقیق‌تر این دو مرحله خواهیم پرداخت.

۳-۴-۱- مرحله برون خط

- خزنده: یک خزنده برای یافتن اسناد GML در سیستم پیشنهادی تعبیه شده است. در مرحله فعلی از این پژوهش خزنده ما فقط بر روی اسناد با قالب GML حساس می‌باشد و نسبت به سایر قالب‌ها بی تفاوت است. به این ترتیب خزنده پس از دریافت خوراک اولیه شروع به کار کرده و اسناد GML را می‌یابد و نشانی وب آنها را در پایگاه داده ذخیره می‌کند.

در بخش تجزیه متن اسناد GML، خزنده پس از یافتن یک سند، متن آنرا تجزیه کرده و بسته به ماهیت هر عنصر، متن تجزیه شده را به تفکیک در جداول جداگانه در پایگاه داده ذخیره می‌کند.

۲- پایگاه داده: پایگاه داده دو وظیفه مهم ذخیره و شاخص‌گذاری اسناد را کنترل می‌کند. پس از اینکه خزنده اسناد GML را یافت، بسته به ماهیت هر المان در متن GML، اطلاعات مختلف را در جدول‌های جداگانه در پایگاه داده ذخیره می‌کند. به عنوان مثال، ابتدا خزنده نشانی وب اسناد GML را در یک جدول خاص ذخیره می‌کند. سپس پس از تجزیه متن GML، برچسب‌های مکانی مثل نوع عارضه، مختصات و ... را در جدول مخصوص به اطلاعات مکانی و برچسب‌های غیرمکانی مثل ویژگی‌ها را در جدول مخصوص به اطلاعات غیرمکانی ذخیره می‌کند.

۳-۴-۲- مرحله بر خط

- واسط کاربری: واسط کاربری تعامل بین کاربر و سیستم را فراهم می‌کند. در سیستم پیشنهادی واسط کاربری مشتمل بر دو روش جستجو می‌باشد. در روش مبتنی بر کلیدواژه کاربران کلیدواژه‌های خود را در واسط کاربری وارد می‌کند و پس از تکمیل فرآیند جستجو، واسط کاربری نتایج را به آن‌ها نمایش می‌دهد.

در روش مختصات مبنا کاربر می‌تواند مختصات نقطه مورد نظر خود را در دو حالت طول و عرض جغرافیایی یا مختصات سیستم تصویر مرکاتور معکوس جهانی[1] (UTM) وارد کند. در حالت UTM کاربر شماره منطقه[2] را در واسط کاربری وارد می‌کند. همچنین در روش مختصات مبنا کاربر موظف به وارد کردن یک حد آستانه براساس متر است. در صورت وارد نکردن حد آستانه، سیستم برای جلوگیری از بار محاسباتی بیش از اندازه به صورت پیش فرض یک حد آستانه برای فرآیند جستجو در نظر می‌گیرد، چرا که در غیر اینصورت تمامی اسناد موجود در پایگاه داده به کاربر برگردانده می‌شوند.

۲- الگوریتم رتبه‌دهی: الگوریتم رتبه‌دهی در دو روش جستجوی مختلف به صورت زیر عمل می‌کند:

در روش کلیدواژه مبنا رتبه‌دهی بر اساس شباهت بین کلیدواژه هدف و کلمات موجود در اسناد انجام می‌پذیرد؛ به این صورت که اسنادی که در متن خود دارای کلمات مشابه بیشتری با کلیدواژه هدف هستند در رتبه‌ای بالاتر در خروجی قرار می‌گیرند.

در روش مبتنی بر مختصات رتبه‌دهی بر مبنای فاصله بین نقطه هدف و عوارض موجود در اسناد صورت می‌گیرد؛ به این صورت که اسنادی که شامل عارضه‌ای با فاصله کمتر نسبت به نقطه هدف هستند حائز رتبه بالاتری خواهند شد. در مورد عوارض چندضلعی در صورتی که نقطه هدف درون مستطیل محاط چندضلعی قرار گیرد، برای رتبه‌دهی عارضه‌ای رتبه بالاتری دریافت می‌کند که فاصله مرکز ثقل آن تا نقطه هدف کمتر باشد.

۴- پیاده سازی

پیاده سازی سیستم پیشنهادی در زبان برنامه نویسی جاوا انجام شده است. پیاده‌سازی در دو مرحله برون خط و برخط به صورت جداگانه در ادامه مورد بررسی قرار خواهد گرفت. برای خزنده در این سیستم از یک خزنده متن باز به نام کراولر۴جی که در جاوا نوشته شده است و از پایگاه داده مای‌اس‌کیوال در بخش پایگاه داده این سیستم استفاده شده است.

۴-۱- مرحله برون خط

۱- طراحی خزنده: ما برای این سیستم از یک خزنده متن‌باز به‌نام کراولر۴جی استفاده کردیم و با ایجاد تغییراتی در برنامه این خزنده آن را موظف کردیم تا فقط برروی اسناد با قالب GML حساس باشد. به این ترتیب خزنده پس از دریافت خوراک اولیه شروع به کار کرده و اسناد GML را می‌یابد و نشانی‌های وب آنها را در پایگاه داده ذخیره می‌کند. ما این سیستم را به صورت آزمایشی بر روی یک کارساز کاربردی به نام تامکت[3] به عنوان یک شبیه سازی از فضای وب پیاده‌سازی کرده‌ایم. به این ترتیب خزنده در نسخه آزمایشی به‌جای جستجو در صفحات وب با دریافت صفحات موجود در کارساز کاربری که به صورت مجازی آدرس گرفته‌اند به عنوان خوراک اولیه وارد لایه‌های مختلف آن شده و اسناد GML موجود در پیوندهای آن‌ها را یافته و در پایگاه داده ذخیره می‌کند.

در بخش تجزیه متن اسناد GML، خزنده پس از یافتن یک سند، متن آنرا تجزیه کرده و بسته به ماهیت هر عنصر، متن تجزیه شده را به تفکیک در جداول جداگانه در پایگاه داده ذخیره می‌کند.

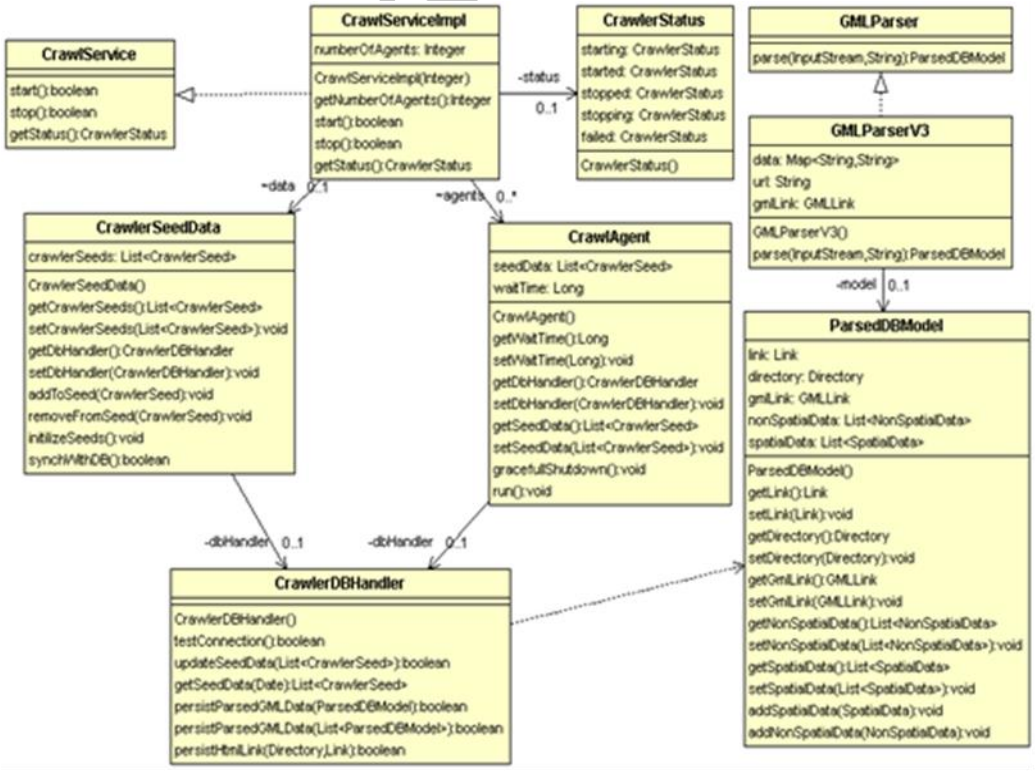
شکل۷ دیاگرام زبان مدلسازی متحد[4] (UML) طراحی شده برای خزنده را در سیستم پیشنهادی نمایش می‌دهد.

در این شکل CrawlerService واسط[5] اصلی خزنده می‌باشد که امکان توقف و اجرای آن را فراهم می‌کند. کلاس CrawlerServiceImpl وظیفه پیاده‌سازی خزنده را بر عهده دارد و به گرفتن خوراک اولیه و تقسیم این خوراک بین مجموعه‌ای از عامل‌ها می‌پردازد. کلاس CrawlAgent وظیفه یافتن اسناد GML و به روز رسانی خوراک اولیه را بر عهده دارد. واسط GMLParser واسط تجزیه اسناد GML می‌باشد و کلاس GMLParserV3 به ترتیب وظیفه تجزیه کردن نسخه‌های ۲ و ۳ اسناد GML را بر عهده دارند. کلاس ParsedDBModel مدل تجزیه شده اسناد GML می‌باشد که در جدوال مختلف پایگاه داده ذخیره می‌شوند. در نهایت کلاس CrawlerDBHandler وظیفه ایجاد سهولت در برقراری

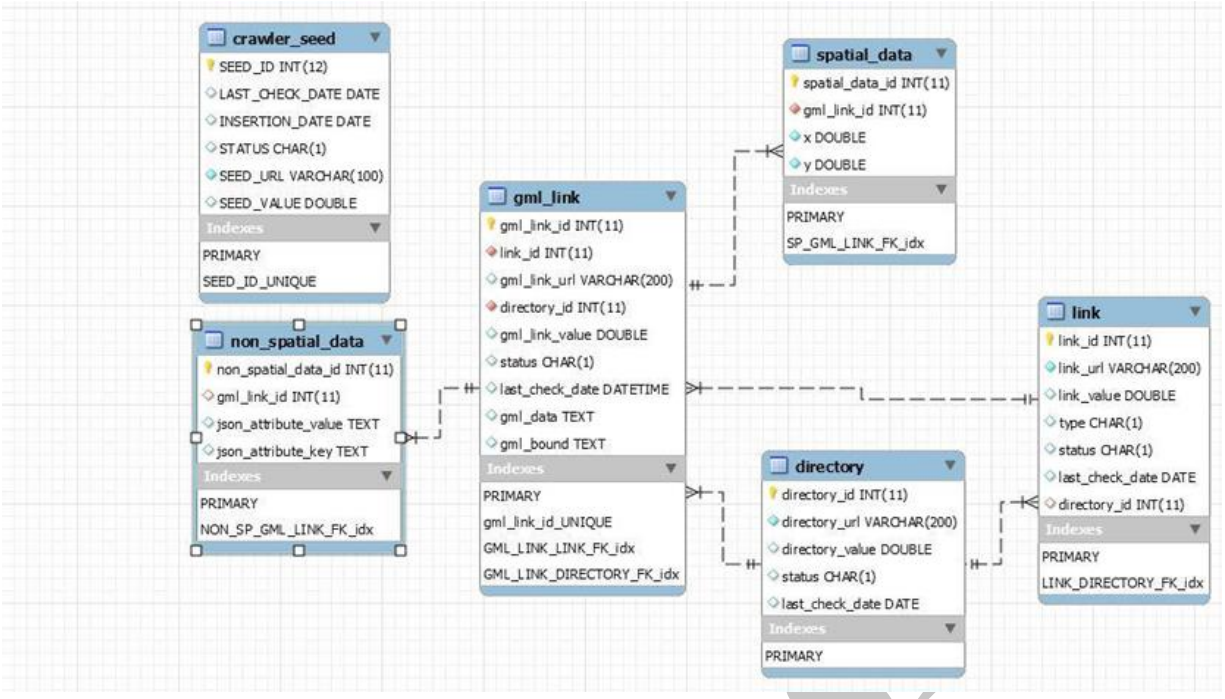
[1] Universal Transvers Mercator
[2] zone
[3] Tomcat
[4] Unified Modeling Language
[5] Interface

ارتباط با پایگاه داده را بر عهده دارد و عامل‌ها با استفاده از این کلاس با پایگاه داده ارتباط برقرار می‌کنند.

۲- طراحی پایگاه داده: همانطور که پیش‌تر اشاره شد پایگاه داده دو وظیفه مهم ذخیره و شاخص‌گذاری اسناد را کنترل می‌کند. پس از اینکه خزنده اسناد GML را یافت، بسته به ماهیت هر المان در متن GML، اطلاعات مختلف را در جدول‌های جداگانه در پایگاه داده ذخیره می‌کند. به عنوان مثال، ابتدا خزنده نشانی وب اسناد GML را در یک جدول خاص ذخیره می‌کند. سپس پس از تجزیه متن GML، برچسب‌های مکانی مثل نوع عارضه، مختصات و ... را در جدول مخصوص به اطلاعات مکانی و برچسب‌های غیرمکانی مثل ویژگی‌ها را در جدول مخصوص به اطلاعات غیرمکانی ذخیره می‌کند. ضمن اینکه این دو جدول با هم ارتباط یک به یک دارند. یعنی هر ردیف از جدول اطلاعات مکانی با یک و فقط یک ردیف از جدول غیرمکانی ارتباط دارد. مشکلی که در اینجا با آن مواجه می‌شویم این است که یک سند GML می‌تواند چندین ویژگی مکانی و غیرمکانی را برای یک عارضه شامل شود، مضاف براینکه یک سند GML می‌تواند چندین عارضه را در خود نگه دارد. راه حلی که برای این مشکل پیشنهاد داده‌ایم استفاده از قالب جیسون(JSON) است. به این صورت که برای هر عارضه تمام اطلاعات مکانی در یک فایل جیسون ذخیره می‌شوند و تمام ویژگی‌های غیرمکانی به طور مشابه در یک فایل جیسون مجزا ذخیره می‌شوند. به این ترتیب هر عارضه دو فایل جیسون دارد که یکی اطلاعات مکانی و دیگری اطلاعات غیرمکانی را ذخیره می‌کند و هرکدام از این فایل‌ها در جداول مخصوص به خود ذخیره می‌شوند. به این ترتیب مشکل ارتباط یک به یک دو جدول اطلاعات مکانی و غیرمکانی حل می‌شود. همچنین این دو جدول یک ارتباط یک به چند با جدول مربوط به ذخیره نشانی‌های وب دارند. به این صورت که هر رکورد اطلاعاتی در جدول مذکور می‌تواند با چند رکورد از جداول مکانی و غیرمکانی در ارتباط باشد، چرا که همانطور که اشاره شد هر فایل GML می‌تواند چندین عارضه را شامل شود. علاوه بر این سه جدول، در نسخه آزمایشی یک جدول هم برای نگه‌داری مسیری که اسناد GML بر روی سیستم ذخیره شده‌اند وجود دارد که در نسخه نهایی از آنجاییکه که اسناد GML بر روی وب قرار دارند این جدول حذف می‌گردد. جدول مورد اشاره یک ارتباط یک به یک با جدول نشانی‌های وب دارد. شکل۸ نمودار ER پایگاه داده را نمایش می‌دهد.



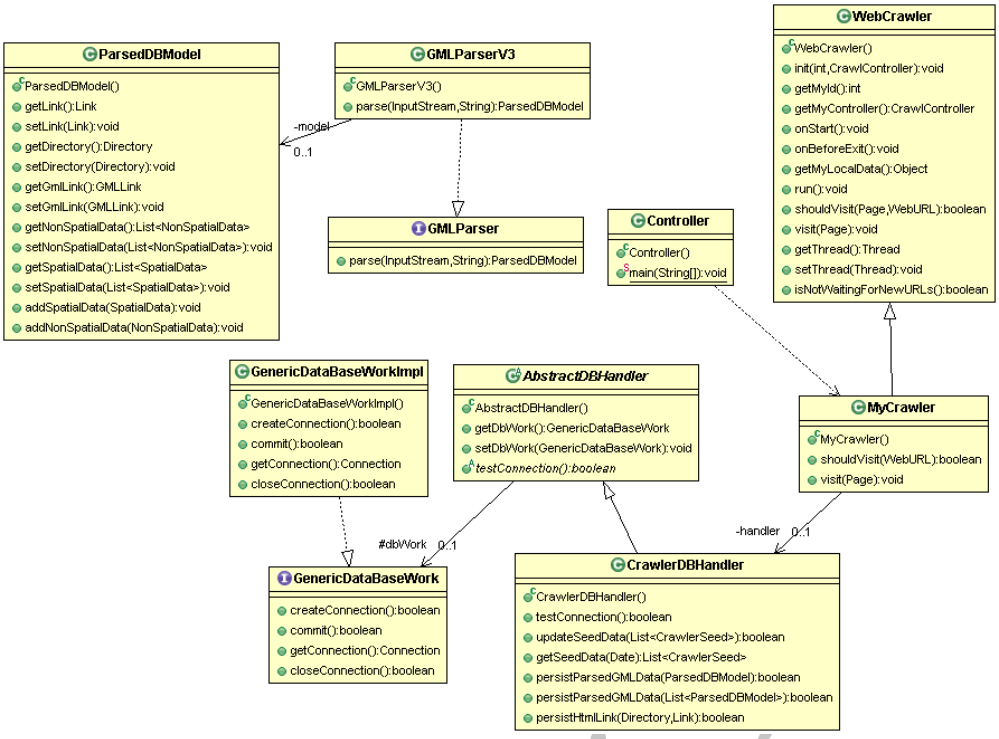
شکل۷- نمودار کلاس‌ها و توابع خزنده UML



شکل۸- دیاگرام ER پایگاه داده

همانطور که در شکل مشخص است اطلاعات مختلف در جداول مخصوص به خود در پایگاه داده ذخیره می‌شوند. اولین نوع اطلاعات داده‌های مکانی موجود در اسناد GML می‌باشد. این داده‌ها که متشکل از مختصات عوارض می‌باشند در جدول spatial_data ذخیره می‌شوند. داده‌های غیر مکانی از دیگر اطلاعات موجود در یک سند GML می‌باشند. این داده‌ها برای هر سند در قالب‌های JSON ذخیره شده و در جدول non_spatial-data ذخیره می‌شوند. جدول link وظیفه ذخیره URL هر سند GML را بر عهده دارد. همچنین در نسخه آزمایشی از آنجایی که اسناد GML در حافظه رایانه ذخیره شده‌اند یک جدول تحت عنوان directory برای ذخیره مسیر این اسناد در حافظه طراحی شده است.

همچنین شکل ۹ نمودار UML مربوط به مرحله برون خط را نمایش می‌دهد. در این شکل، کلاس Controller کلاس اجرا کننده فاز برون خط می‌باشد و نقطه شروع برنامه به حساب می‌آید. کلاس WebCrawler خزنده را در حالت کلی نمایش می‌دهد و کلاس MyCrawler خزنده‌ای است که از کلاس WebCrawler ساخته شده است و فرزند آن محسوب می‌شود. در واقع کلاس WebCeawler کلاس مربوط به خزنده Crawler4j می‌باشد که کلاس MyCrawler از روی آن ساخته شده است این خزنده شامل قسمتی است که با کلاس‌های پایگاه داده در ارتباط است. کلاس CrawlerDBHandler برجسته‌ترین کلاس در این قسمت می‌باشد. این کلاس فعالیت‌هایی نظیر ذخیره اطلاعات تجزیه شده در پایگاه داده را انجام می‌دهد. کلاس CrawlerDBHandler از یک کلاس پدر به نام AbstractDBHandler که به صورت انتزاعی کار با پایگاه داده را نشان می‌دهد به وجود آمده است. کلاس GenericDataBaseWorkImpl نیز پیاده‌سازی فعالیت‌های ساده کار با پایگاه داده را انجام می‌دهد. از طرفی کلاس GMLParser وظیفه تبدیل اطلاعات رشته‌ای که از اسناد GML خوانده شده است به اشیاء قابل تفسیر را برعهده دارد. در واقع این کلاس داده رشته‌ای را به مدل داده‌ای موجود در پایگاه داده تبدیل می‌کند.



شکل۹- دیاگرام UML مرحله برون خط

۴-۲- مرحله بر خط

۱- طراحی واسط کاربری: واسط کاربری در این سیستم به گونه‌ای طراحی شده است که کاربر بتواند جستجوی خود را به روش دلخواه انجام دهد.

واسط کاربری از دو شیوه جستجو پشتیبانی می‌کند. در روش کلیدواژه‌مبنا کاربر کلیدواژه‌های خودر را به سیستم ارسال می‌کند و پس از طی فرآیند جستجو، تنایج به کاربر برگردانده می‌شوند.

در روش مختصات‌مبنا نیز کاربر می‌تواند مختصات یک نقطه هدف را به سیستم ارسال کند. همچنین در این روش جستجو کاربر می‌تواند مختصات نقطه مورد نظر خود را در دو حالت طول و عرض جغرافیایی یا مختصات سیستم تصویر UTM در واسط کاربری وارد کند.

در واسط کاربری همچنین کاربر قادر است یک گستره از منطقه مورد نظر خود را به صورت نقشه مشاهده کند. برای پیاده‌سازی این بخش از سیستم از امکاناتی که موتور جستجوی Google برای نمایش نقشه مهیا کرده، استفاده شده است برای نمایش نقشه مهیا کرده، استفاده شده است در بخش نتایج نیز اسناد موجود در پایگاه داده که مرتبط با جستجو کاربر می‌باشد نمایش داده می‌شود. همچنین کاربر می‌تواند در واسط کاربری یک حد آستانه وارد کند تا تنها عوارضی در خروجی فهرست شوند که فاصله آنها از این حد آستانه کمتر باشد. علت استفاده از حد آستانه این است که در صورت عدم وجود آن تمامی اسناد موجود در پایگاه داده در خروجی به نمایش در می‌آیند که یقیناً کاربر نیازی به تمامی این اسناد ندارد و همچنین بار محاسباتی زیادی به سیستم تحمیل می‌شود و عملکرد سیستم را تحت تاثیر قرار خواهد داد. ترتیب نمایش این اسناد در خروجی براساس میزان شباهت هر سند با جستجوی کاربر می‌باشد که در بخش رتبه‌دهی سیستم مشخص می‌شود.

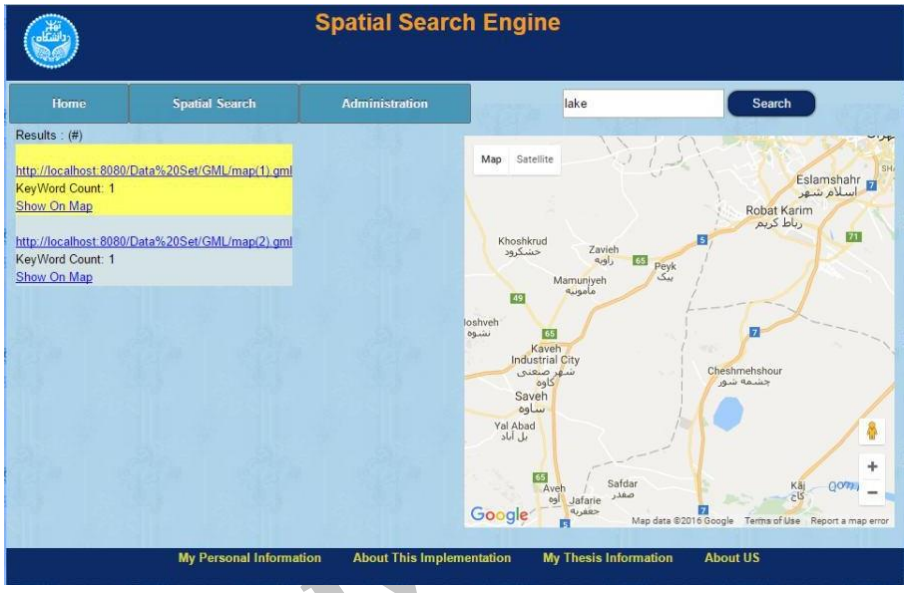
۲- مکانیزم رتبه‌دهی: همانطور که پیش‌تر اشاره شد فرآیند رتبه‌دهی برای هر شیوه جستجو به صورت جداگانه عمل می‌کند.. در روش کلیدمبنا شباهت بین کلیدواژه جستجو شده و اسناد موجود اساس مکانیزم رتبه‌دهی رو تشکیل می‌دهد. به این ترتیب که کلیدواژه مذکور در میان اسناد موجود در پایگاه داده جستجو می‌شود و اسنادی که کلیدواژه در انها عیناً بیشتر تکرار شود رتبه بالاتری را از آن خود می‌کند.

در روش بر مبنای مختصات ابتدا فاصله بین نقطه ارسال شده توسط کاربر و عوارض موجود در اسناد ذخیره شده در پایگاه داده محاسبه می‌شود. سپس اسنادی که حداقل یکی از عوارض موجود در آن فاصله کمتری نسبت به نقطه هدف دارند حائز رتبه بالاتری می‌شوند و در نتایج قبل از سایر اسناد به کاربر نمایش داده می‌شوند. لازم به

ذکر است تمامی اسنادی که فاصله یکی از عوارض انها از حد استانه تعیین شده توسط کاربر کمتر است در خروجی فهرست می‌شوند. ضمناً در صورتی که حد آستانه‌ای توسط کاربر مشخص نشود سیستم به صورت پیش فرض حد آستانه را ۱۰۰۰ متر در نظر می‌گیرد. همچنین در مورد عوارض چندضلعی به جای حد آستانه از مستطیل محاط استفاده می‌شود. به این صورت که اسنادی که نقطه هدف درون مستطیل محاط یکی از عوارض موجود در آنها قرار بگیرد در خروجی به کاربر نمایش داده می‌شوند. در مورد رتبه‌دهی اینگونه اسناد فاصله نقطه هدف تا مرکز ثقل محاسبه می‌گردد، به طوری که فاصله کمتر رتبه بالاتری را دریافت می‌کند.

۴-۳- ارائه نتایج

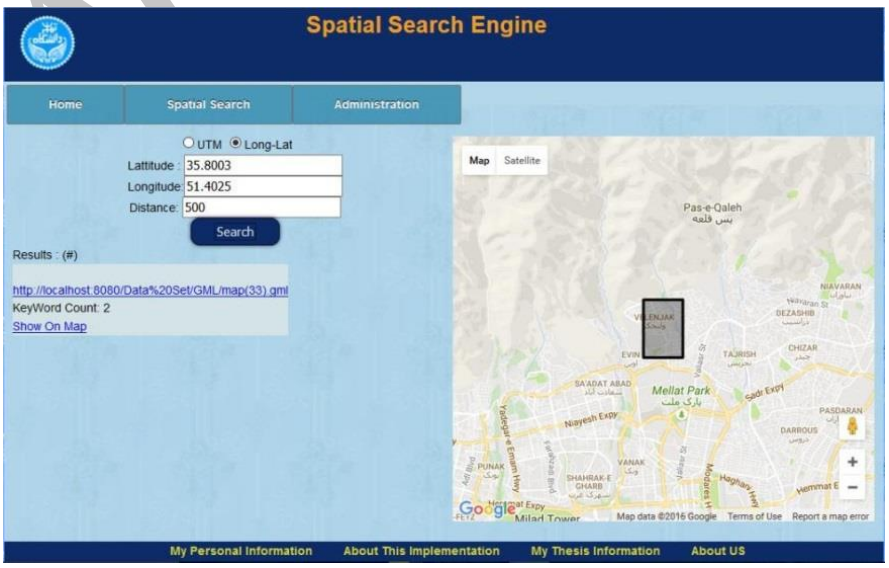
شکل۱۰ نمونه‌ای از خروجی سیستم را برای یک جستجوی بر مبنای کلیدواژه برای کلیدواژه جستجو شده lake نشان می‌دهد.



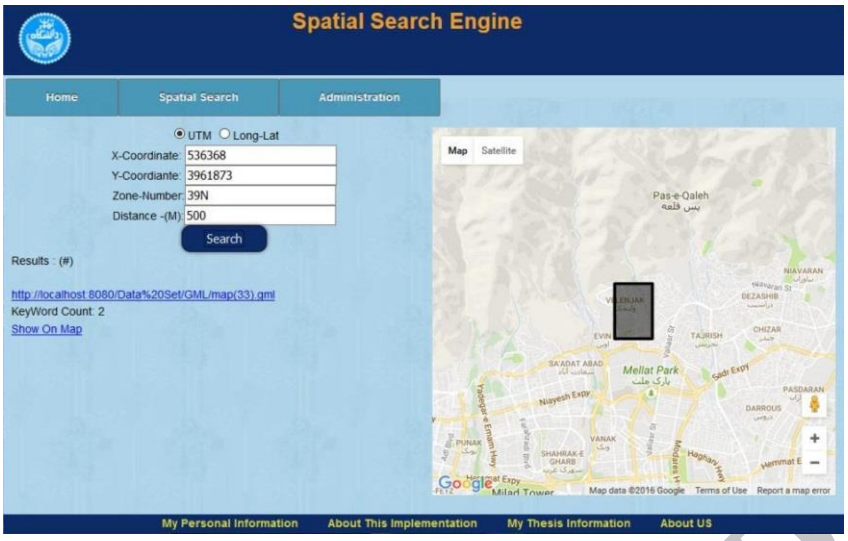
شکل۱۰- خروجی سیستم برای یک جستجوی بر مبنای کلیدواژه

همانطور که در شکل مشخص است خروجی سیستم برای کلیدواژه جستجو شده شامل دو سند GML موجود در پایگاه داده است که شامل کلیدواژه lake می‌باشند.

همچنین شکل‌های ۱۱ و ۱۲ خروجی‌های سیستم را به ترتیب برای یک جستجو براساس طول و عرض جغرافیایی و سیستم تصویر UTM نمایش می‌دهند.



شکل۱۱- خروجی سیستم برای یک جستجو براساس طول و عرض جغرافیایی



شکل ۱۲- خروجی سیستم برای یک جستجو براساس مختصات سیستم تصویر UTM

۴-۳- اعتبارسنجی

برای ارزیابی و اعتبارسنجی سیستم پیشنهادی از مقایسه آن با ۴ موتور جستجوی مکانی معروف براساس ۸ پارامتر موثر در یک موتور جستجو استفاده شده است. پارامترهای مورد بررسی عبارتند از:

- نمایش نقشه
- جستجو براساس کلیدواژه
- جستجو براساس مختصات
- جستجو براساس فاصله
- رتبه‌دهی براساس ویژگی‌های کاربر
- رتبه‌دهی براساس فاصله
- قابلیت تجزیه اسناد مکانی
- استفاده از پایگاه داده برای ذخیره اطلاعات

همچنین برای مقایسه از چهار موتور جستجوی SPIRIT، SOLR، Elasticsearch و Google استفاده شده است. جدول ۱ هرکدام از معیارهای فوق را برای موتورهای جستجوی مذکور مورد مقایسه قرار داده است.

جدول۱- مقایسه موتورهای جستجوی مکانی براساس معیارهای مختلف

	قابلیت تجزیه اسناد مکانی	جستجو بر اساس مختصات	جستجو براساس فاصله	رتبه‌دهی براساس فاصله	رتبه‌دهی براساس ویژگی‌های کاربر	جستجو براساس کلیدواژه	نمایش نقشه	استفاده از پایگاه داده
سیستم پیشنهادی	✓	✓	✓	✓	×	✓	✓	✓
SOLR	×	✓	✓	✓	×	×	✓	×
Elasticsearch	×	✓	✓	✓	×	×	✓	×
SPIRIT	×	×	✓	✓	×	✓	✓	✓
Google Map	×	✓	×	×	✓	✓	✓	✓

مشاهده می‌شود که هیچکدام از این ۴ موتور جستجوی معروف به تجزیه اسناد مکانی توجهی نداشته‌اند و با توجه به اهمیت اسناد مکانی و اطلاعات نهفته در آن‌ها، این مسئله گواهی است بر این ادعا که سیستم حاضر گام موثری در بازیابی اطلاعات مکانی در موتورهای جستجو برداشته است. لازم به توضیح است که از آنجایی که کار ارائه شده در این مقاله صرفاً جنبه پژوهشی دارد، بدون شک مقایسه این سیستم با موتورهای قدرتمندی نظیر Google از لحاظ سرعت عملکرد قیاس مع الفارغ است.

۵- نتیجه‌گیری و پیشنهادات

علی‌رغم وجود حجم قابل توجه اسناد و اطلاعات مکانی در وب، به دلیل گستردگی دنیای وب به سختی می‌توان این اسناد و اطلاعات را در حجم عظیم اطلاعات

دیگر یافت. اکثر موتورهای جستجوی مکانی امروزی بر روی نمایش نقشه تمرکز دارند و داده مکانی کمتر مورد توجه قرار گرفته است. این درحالیست که امروزه مهندسین و متخصصین زیادی به اطلاعات مکانی خام نیاز دارند و به دنبال راهی برای یافتن این اطلاعات در دنیای وسیع وب هستند. این مقاله نشان می‌دهد که چگونه افزودن توان استخراج اطلاعات مکانی و غیر مکانی از اسناد GML می‌تواند پاسخگوی بسیاری از پرسش‌ها باشد که در حال حاضر مورد مسامحه قرار گرفته‌اند. سیستم طراحی شده به عنوان یک موتور جستجوی مکانی امکان جستجو در بین اسناد GML را برای کاربر فراهم کرده و به این ترتیب گام مؤثری در بهبود کارایی موتورهای جستجوی مکانی برداشته است. از آنجایکه اسناد GML شامل اطلاعات صریح مکانی در کنار اطلاعات غیر مکانی است، مهمترین مزیت این سیستم در مقایسه با سایر موتورهای جستجوی مکانی نگرش یکپارچه به اطلاعات مکانی و غیرمکانی می‌باشد.

امروزه اطلاعات مکانی در قالب‌های مختلفی مثل کی ام ال[۱]، ژئوجیسون[۲]، توپوجیسون[۳] و ... ذخیره می‌شوند. نتایج به دست آمده از سیستم پیشنهادی با ایجاد تغییراتی اندک در خزنده به راحتی قابل تعمیم به سایر قالب‌های مکانی می‌باشد. ایجاد یک سیستم یکپارچه که قابلیت شناسایی، درک، تحلیل و تجزیه اینگونه قالب‌ها و همچنین تبدیل انواع قالب‌های مکانی به یکدیگر را داشته باشد، می‌تواند به عنوان ادامه مسیر پژوهش حاضر مطرح شود. همچنین بهینه‌سازی اجزای سیستم معرفی شده از جمله خزنده جهت بهبود کارایی سیستم از دیگر اهداف مدنظر برای آینده می‌باشد.

مراجع

[1] Jones CB, Abdelmoty AI, Finch D, Fu G, Vaid S. The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. Geographic Information Science: Springer; 2004. p. 125-39.

[2] Spink A, Jansen BJ. A study of web search trends. Webology. 2004;1(2):4.

[3] Asadi S, Chang C-Y, Zhou X, Diederich J. Searching the world wide web for local services and facilities: A review on the patterns of location-based queries. Advances in Web-Age Information Management: Springer; 2005. p. 91-101.

[4] Zhang J ,Gruenwald L, editors. A GML-based open architecture for building a geographical information search engine over the Internet. wise; 2001: IEEE.

[5] Sanderson M, Kohler J, editors. Analyzing geographic queries. SIGIR Workshop on Geographic Information Retrieval; 2004.

[6] Bone C, Ager A, Bunzel K, Tierney L. A geospatial search engine for discovering multi-format geospatial data across the web. International Journal of Digital Earth. 2014:1-16.

[7] Greenwood J, Whiteside A. OGC Web Services Common Standard. Citeseer; 2010.

[8] Asadi S, Zhou X, Jamali HR, Mofrad HV. Location-based search engines tasks and capabilities: A comparative study. Webology. 2007;4(4).

[9] Walter V, Luo F, Fritsch D. Automatic Map Retrieval and Map Interpretation in the Internet. Advances in Spatial Data Handling: Springer; 2013. p. 209-21.

[10] Zhao J, Jin P, Zhang Q, Wen R. Exploiting location information for web search. Computers in Human Behavior. 2014;30:378-88.

[11] Amitay E, Har'El N, Sivan R, Soffer A, editors. Web-a-where: geotagging web content. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval; 2004: ACM

۱ KML
۲ GeoJSON
۳ TopoJSON