

استخراج نام‌های جغرافیایی از داده‌های وب با استفاده از تکنیک وب‌کاوی

مهدی تیموری پابندی^۱، محمد رضا ملک^۲، علی اصغر آل‌شیخ^۳

۱- دانشجوی کارشناسی ارشد، گروه سیستم‌های اطلاعات مکانی دانشکده ژئودزی و ژئوماتیک، دانشگاه خواجه

نصیرالدین طوسی

teimurimehdi@gmail.com

۲- استادیار، گروه سیستم‌های اطلاعات مکانی دانشکده ژئودزی و ژئوماتیک، دانشگاه خواجه نصیرالدین طوسی

mrmalek@kntu.ac.ir

۳- استادیار، گروه سیستم‌های اطلاعات مکانی دانشکده ژئودزی و ژئوماتیک، دانشگاه خواجه نصیرالدین طوسی

alesheikh@kntu.ac.ir

چکیده

رشد سریع و روزافزون وب پتانسیل بالایی برای استخراج اطلاعات و دانش مفید از آن ایجاد کرده است. در این پژوهش ابتدا وب‌کاوی مورد بررسی قرار می‌گیرد، وب‌کاوی دانشی است که از تکنیک‌های داده‌کاوی به منظور استخراج اطلاعات و دانش مفید از داده‌های وب استفاده می‌کند. سپس کاربردهای آن در حوزه اطلاعات مکانی به صورت مختصر مرور می‌شود، و فعالیت‌های مکانی انجام شده در این حوزه بررسی می‌شوند، اغلب این فعالیت‌ها نوظهور و عمدتاً مبتنی بر کاوش متون است. می‌دانیم ایجاد اطلس‌های جغرافیایی یک فرآیند هزینه‌بر و نیازمند کار فراوان است. با شناسایی نام مکان‌ها از اسناد وب می‌توان یک گام به ساخت خودکار اطلس‌ها نزدیک شد. در این پژوهش از تکنیک‌های وب‌کاوی برای استخراج نام‌های پیشین شهرهای ایران استفاده شده است. در نتیجه این پژوهش نام‌های پیشین شهرهای ایران با دقت خوبی از اسناد ویکی‌پدیا استخراج شد. این پژوهش می‌تواند نقطه شروعی برای پژوهش‌های این‌چنینی باشد.

واژه‌های کلیدی: وب، وب‌کاوی، کاوش متون، نام‌های جغرافیایی

۱-۱ تعریف مساله

اختراع وب این امکان را برای کاربران فراهم کرده است که دسترسی فراوانی به اطلاعات داشته باشند، تقریباً هر موضوعی که فکرش را بتوان کرد. یک نفر می‌تواند به اطلاعاتی که توسط دیگر کاربران تولید شده است، دسترسی داشته باشد، این کاربران محدوده وسیعی را شامل می‌شوند، از افراد معمولی گرفته تا شرکت‌های بزرگی که در وب تجارت می‌کنند. بخش بزرگی از این داده‌ها حاوی اطلاعات مکانی است، در نتیجه فرصت بزرگی برای استخراج داده‌های مکانی فراهم می‌شود. هدف این پژوهش ارزیابی استخراج نام‌های مکانی با استفاده از تکنیک‌های داده‌کاوی از داده‌های داوطلبانه فراهم شده توسط کاربران در وب است، به این منظور سایت ویکی‌پدیا استفاده شده است، می‌خواهیم ببینیم آیا امکان استخراج نام مکان‌ها از داده‌های وب وجود دارد یا خیر؟

۲-۱ ضرورت و اهمیت تحقیق

وب فرصت و چالش بی‌نظیری برای داده‌کاوی فراهم می‌کند. دلیل آن را می‌توان در خصوصیات وب جستجو کرد [۱]:

۱- مقدار عظیمی داده بر روی وب وجود دارد که هر روز در حال رشد است. این داده‌ها به راحتی در دسترس‌اند.

۲- پوشش اطلاعاتی وب، وسیع و گوناگون است. هر گونه اطلاعاتی را می‌توان در سطح وب یافت.

۳- داده در فرمت‌های مختلف بر روی وب وجود دارد: جداول ساختاریافته، متون، داده‌های چندرسانه‌ای.

۴- اطلاعات موجود در وب غیرمتجانس است. چندین صفحه وب ممکن است، اطلاعات یکسان یا مشابه را به صورت‌های گوناگون نمایش دهند، که می‌تواند تلفیق اطلاعات را به یک وظیفه چالش برانگیز تبدیل کند.

۵- بسیاری از اطلاعات وب نیمه ساختاریافته‌اند، زیرا از کدهای تودرتوی HTML استفاده می‌کنند. چرا که طراحان وب نیاز دارند، اطلاعات را به روش‌های ساده و معمول نمایش دهند، تا برای مشاهده و فهم انسان ساده باشد.

۶- بسیاری از اطلاعات وب به هم لینک شده‌اند. این ارتباط بین صفحات درون یک سایت و نیز با صفحات سایت‌های دیگر وجود دارد. این لینک‌ها به عنوان ابزار سازماندهی اطلاعات و نیز به عنوان معیاری برای ارزیابی اعتمادپذیری و صحت صفحات لینک شده وب‌سایت‌ها استفاده می‌شوند.

۷- افزونگی زیادی در اطلاعات وب وجود دارد.

۸- وب دارای نویز است. یک صفحه وب به طور معمول شامل مخلوطی از انواع مختلف اطلاعات است. به طور مثال محتوای اصلی، تبلیغات، هشدارهای کپی‌رایت و غیره. برای یک کاربرد خاص تنها بخشی از اطلاعات مفید است و بقیه نویز است.

۹- وب از وب سطحی و وب عمیق تشکیل شده است. وب سطحی از صفحاتی تشکیل شده است که می‌توان به وسیله یک مرورگر معمولی آنها را مرور کرد. وب سطحی همچنین از طریق موتورهای جستجو قابل جستجو است. وب عمیق از پایگاه داده‌هایی تشکیل شده است که تنها از طریق پرسش‌وپاسخ و با استفاده از فرم‌ها قابل دسترس‌اند.

۱۰- وب همچنین درباره سرویس‌هاست. بسیاری از وب‌سایت‌ها و صفحات مردم را قادر می‌سازند تا عملیاتی را با استفاده از پارامترهای ورودی انجام دهند، مانند خرید.

۱۱- وب پویا است. اطلاعات بر روی وب به طور ثابت تغییر می‌کنند، همگامی با تغییرات و مانیتور کردن آنها موضوع مهمی برای بسیاری از کاربردهاست.

از سوی دیگر بخش بزرگی از منابع موجود در وب به اطلاعاتی بر می‌گردد که می‌توان آنها را به عنوان داده‌های مکان مبنا در نظر گرفت. بدین معنی که بیشتر فعالیت‌ها و سرمایه‌گذاری‌ها در یک یا چند نقطه از زمین اتفاق می‌افتند، و منابع بزرگی از داده‌های نقشه‌برداری، تصاویر، نقشه‌ها و گزارش‌هایی وجود دارد که به یک مکان یا موقعیت خاص مرتبط‌اند [۲]. نشان داده شده است که حدود ۱۸ درصد از عبارات پرس‌وجو در موتورهای جستجو حاوی عبارات مکانی است، تقریباً ۲۰ درصد از صفحات وب شامل یک یا تعداد بیشتری از شاخص‌های مکانی است [۳].

همانطور که دیدیم وب یک راه بسیار مناسب، کم هزینه و عمومی برای انتشار، اخذ و در دسترس قرار دادن اطلاعات در اختیار عموم مردم است و به یکی از متداول‌ترین منابع ارتباطی و اطلاعاتی جهان تبدیل گردیده است، این منبع شامل مقادیر بسیاری اطلاعات مکانی است، آشکار است که پتانسیل بالایی برای استخراج دانش و اطلاعات مفید از وب وجود دارد. وب‌کاوی برای استخراج چنین اطلاعاتی فعالیت می‌کند.

وب‌کاوی

وب داری حجم زیادی از داده‌ها به اشکال زیر است [۴]:

- محتوا^۱: داده‌های واقعی در صفحات وب. به عبارت دیگر داده‌هایی که صفحات وب برای رساندن آن‌ها به کاربران ایجاد شده‌اند. محتوا معمولاً ولی نه لزوماً شامل متن و گرافیک است.
- ساختار^۲: داده‌هایی که سازمان محتوا را توصیف می‌کنند. این داده‌ها دو نوع اند، اطلاعات ساختار درون صفحه‌ای شامل چیدمان تگ‌های مختلف ^۳HTML و ^۴XML در درون یک صفحه. و اطلاعات ساختار بین صفحات که توسط فرایوندها^۵ به یکدیگر متصل می‌شوند.
- داده‌های استفاده^۶: داده‌هایی که الگوی استفاده از صفحات وب را توصیف می‌کنند. مانند آدرس ^۷IP، مرجع صفحات و تاریخ و زمان دسترسی.
- نمایه کاربر^۸: داده‌هایی که اطلاعات آماری درباره کاربران وب‌سایت فراهم می‌کنند، مانند اطلاعات ثبت شده در فرم‌ها.

^۱ - Content

^۲ - Structure

^۳ - Hyper Text Markup Language

^۴ - Extensible Markup Language

^۵ - Hyperlinks

^۶ - Usage

^۷ - Internet Protocol

^۸ - User Profile

با توجه به رشد بسیار سریع و روزافزون وب، یافتن اطلاعات موردنیاز از میان این انبوه اطلاعات کار بسیار مشکلی است و ضرورت وجود تکنیک‌هایی برای بررسی و سازماندهی داده‌ها، به منظور اخذ نتیجه بهتر و بازدهی موثرتر را نمایان می‌سازد. وب‌کاوی دانشی است که از تکنیک‌های داده‌کاوی به منظور کشف و استخراج خودکار دانش از اسناد و سرویس‌های وب استفاده می‌نماید. متدهای وب‌کاوی می‌توانند برای یافتن منابع، انتخاب و پیش‌پردازش اطلاعات، استخراج و آنالیز الگوهای دسترسی به صفحات یا سایت‌ها به کار گرفته شوند. می‌توان وب‌کاوی را توسعه اکتشاف دانش^۱ بر روی داده‌های تحت وب دانست [۵].

در ادامه روش‌های وب‌کاوی مورد بررسی قرار می‌گیرد. وب‌کاوی بنا بر نوع داده‌ای که مورد کاوش قرار می‌گیرد، به سه دسته اصلی تقسیم می‌شود که عبارتند از محتوا کاوی وب^۲، ساختار کاوی وب^۳، و استفاده کاوی وب^۴ [۵]. در ادامه هر کدام از این دسته‌ها به طور مختصر مورد بررسی قرار می‌گیرند. در شکل ۱ ساختار درختی این طبقه‌بندی را مشاهده می‌کنید.

۲-۱ محتوا کاوی وب

محتوا کاوی وب، کاربرد تکنیک‌های داده‌کاوی بر روی محتوای انتشار یافته در اینترنت است، که معمولاً به صورت HTML (نیمه ساختار یافته)، متن خالی^۵ (ساختار نیافته) یا XML (ساختار یافته) منتشر گردیده‌اند [۶]. منبع اصلی وب که باید مورد کاوش قرار گیرد، یک صفحه منفرد است، بنابراین این بخش از وب‌کاوی شامل محدوده بسیار وسیعی از اطلاعات مانند داده‌های متنی، تصویری، و صوتی می‌شود که ساختار نیافته‌اند، همچنین می‌تواند شامل اطلاعات ساختاریافته مانند لیست‌ها و جداول باشد. از تکنیک‌های متن‌کاوی^۶ (متن‌کاوی یک زیر دسته از محتوا کاوی وب است که از ساختار وب استفاده نمی‌کند) به صورت گسترده در محتوا کاوی وب استفاده می‌شود. از متن‌کاوی در اکتشاف موضوع، استخراج الگوهای شباهت، خوشه‌بندی اسناد وب و دسته‌بندی صفحات استفاده می‌شود [۷].

^۱ - Knowledge Discovery

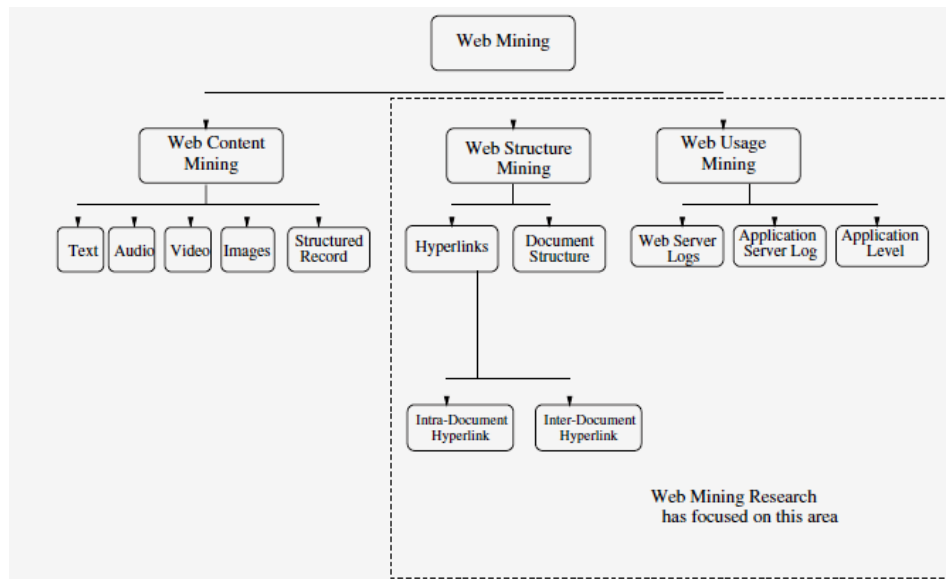
^۲ - Web Content Mining

^۳ - Web Structure Mining

^۴ - Web Usage Mining

^۵ - Plaintext

^۶ - Text Mining



شکل ۱- طبقه‌بندی وب‌کاوی [۷]

۲-۲ ساختار کاوی وب

ساختار کاوی وب بر روی ساختار فرایوندهای وب انجام می‌شود و به کشف مدل حاکم بر ساختار فرایوندهای وب می‌پردازد. این ساختار گرافی می‌تواند اطلاعاتی درباره رتبه‌بندی صفحات وب فراهم کند و نتایج جستجو را از طریق فیلترینگ بهبود بخشد [۶]. یک بخش مهم از کاربرد این کاوش مشخص کردن رابطه نسبی صفحاتی است که وقتی بر اساس محتوایشان به تنهایی آنالیز می‌شوند، به یک اندازه مناسب به نظر می‌رسند. منبع اصلی مورد کاوش مجموعه‌ای از صفحات است از یک وب‌سایت گرفته تا همه وب. ساختار یک گراف وب معمول از صفحات وب به عنوان نودها و فرایوندها به عنوان یال‌های متصل کننده صفحات تشکیل شده است. ساختار کاوی وب فرایند اکتشاف اطلاعات ساختاری از وب است.

۳-۲ استفاده کاوی وب

استفاده کاوی وب نتایج حاصل از تبادلات کاربر با یک وب‌سرور^۱ را تحلیل می‌کند، به این تبادلات داده‌های استفاده گفته می‌شود، که شامل *web server logs* کلیک‌های انجام شده، تراکنش‌های پایگاه‌های داده در یک وب‌سایت یا در گروهی از وب‌سایت‌ها، دانلودها، خریدها و نیز زمان انجام هر رویداد می‌شوند. استفاده کاوی وب موجب ایجاد نگرانی‌هایی در مورد حریم خصوصی کاربران می‌گردد [۶]. از سوی دیگر می‌توان استفاده کاوی وب را کاربرد تکنیک‌های داده کاوی، به منظور اکتشاف الگوهای استفاده^۲ جالب با استفاده از داده‌های وب دانست، که به فهم بهتر نیازهای برنامه‌های کاربردی تحت وب می‌انجامد و این اجازه را به آن‌ها می‌دهد تا خدمات بهتری ارائه دهند [۴].

^۱ - Web Server

^۲ - Usage Patterns

هدف از انجام استفاده‌کاوی می‌تواند ساخت یک مدل از رفتار کاربران یا ساخت یک سیستم تطبیق‌پذیر^۱ باشد. هدف سیستم‌های تطبیق‌پذیر بهبود برخی از جنبه‌های رفتار سیستم است. می‌توان بسیاری از جنبه‌های رفتار کاربران را مدل کرد. به طور مثال تبادلات با سیستم، علایق، دانش، رفتار جغرافیایی و همچنین ترکیب این رفتارها با یکدیگر. مدل‌سازی ترجیحات نیازمند اطلاعاتی درباره ترجیحات کاربران درباره موضوعات مختلف است. دستیابی به این اطلاعات مشکل است، زیرا کاربران به ندرت آمادگی دارند یک موضوع را ارزیابی کنند و آن را در اختیار طراحان قرار دهند [۸].

۲-۴ مقایسه روش‌های وب‌کاوی

در جدول ۱ مقایسه مختصری از انواع روش‌های وب‌کاوی را مشاهده می‌کنید. محتواکاوی وب از دو منظر پایگاه‌داده (DB) و همین‌طور بازیابی اطلاعات (IR) مورد بررسی قرار گرفته است [۵].
جدول ۱- مقایسه انواع روش‌های وب‌کاوی [۵]

	Web Mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	Information Retrieval View	Database View		
View Of Data	-Unstructured -Semi Structured	-Semi Structured Web site as DB	-Links Structure	-Interactivity
Main Data	- Text documents - Hypertext documents	- Hypertext documents	- Links Structure	- Server logs - Browser Logs
Representation	- Bags of Words, n-grams - Terms, phrases - Concepts or ontology - Relational	- Edge-labeled graph(OEM) - Relational	- Graph	- Relational table - Graph
Method	- TFIDF and variants - Machine Learning -Statistical(Including NLP)	- Proprietary algorithms - ILP - (Modified) association rules	- Proprietary algorithms	- Machine Learning -Statistical(Including NLP) - (Modified) association rules
Application Categories	- Categorization - Clustering - Finding extraction rules - Finding patterns in text	- Finding frequent substructures - Web site schema discovery	- Categorization - Clustering	- Site construction, adaption, and management - Marketing -User modeling

¹ - Adaptive System

۲-۵ بررسی کاربردهای وب‌کاوی در حوزه داده‌های مکانی

استخراج اطلاعات مکانی از منابع گوناگون وب، برای کاربردهای مختلف مهم است. مقدار فراوانی از اطلاعات مکانی در سطح وب وجود دارد، این مقدار مدام در حال رشد است، ولی کشف داده‌ها و تبدیل آن‌ها به فرمت مناسب یک چالش جدی است. همان‌طور که در [۹] بحث شده است، افزایش فزاینده استفاده کاربران از خدمات سرویس‌هایی مانند، ^۱Google Earth، ^۲Wikimapia، ^۳Openstreetmap، امکان ژئوتگ^۴ تصاویر در ^۵Flickr و بسیاری از سرویس‌های دیگر، یک مخزن بزرگ از اطلاعات جغرافیایی داوطلبانه^۶ ایجاد می‌کند. این اطلاعات می‌توانند محدوده وسیعی از نوشته‌ها، تصاویر، صداها، ویدئوها و مدل‌های سه‌بعدی ژئوتگ شده گرفته تا حتی نام‌گذاری و تعیین محدوده خیابان‌ها، محلات، نقاط و مسیرهای دیدنی، فروشگاه‌ها و غیره توسط مردم عادی را شامل شوند.

از تکنیک‌های وب‌کاوی می‌توان برای استخراج و تحلیل این حجم بزرگ اطلاعات استفاده کرد. به طور مثال Pultar و همکاران [۱۰] یک عامل مکانی خزنده در وب^۷ ایجاد کرده‌اند که می‌تواند داده‌های صفحات وب را بخواند، لینک‌ها را دنبال کند و اطلاعات مکانی مورد استفاده در GIS را جمع‌آوری کرده و آنالیزهای داده‌کاوی مکانی^۸ را بر روی آن‌ها انجام دهد. این پروژه ^۹GEDMWA نامیده می‌شود. هدف اصلی این پروژه ایجاد یک عامل مکانی به منظور تعیین موقعیت اطلاعات جغرافیایی داوطلبانه در وب، و نیز تولید داده مکانی معتبر (KML, Shapefile) از این داده‌های اکتشاف شده است. چنین عامل مکانی برای انجمن‌های کاربری گوناگون در سطح جهان می‌تواند سودمند باشد. به طور مثال برای یک گروه از کوهنوردان به منظور یافتن مناطق کوهنوردی جدید. عده زیادی از کوهنوردان اطلاعات خود را در سطح وب همراه با مختصات جغرافیایی و توضیحاتی در مورد مسیرهای کوهنوردی منتشر می‌کنند، از این اطلاعات می‌توان برای کشف مناطق جدید استفاده کرد. در شکل ۲ شمای کلی این پروژه را مشاهده می‌کنید.

^۱ - <http://www.earth.google.com>

^۲ - <http://www.wikimapia.org>

^۳ - <http://www.openstreetmap.org>

^۴ - Geotag

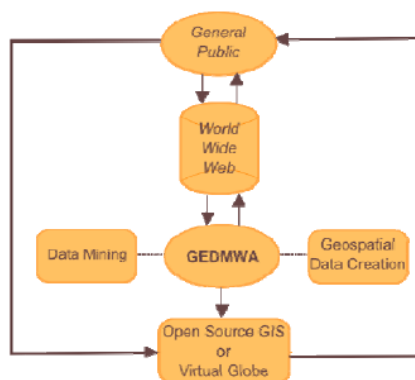
^۵ - <http://www.flickr.com>

^۶ - Volunteered Geographic Information(VGI)

^۷ - Web Crawling Geospatial Agent

^۸ - Spatial Data Mining

^۹ - Geospatial Exploratory Data Mining Web Agent

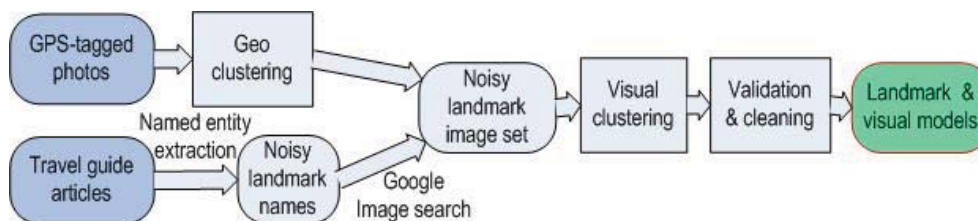


شکل ۲- شمای کلی پروژه GEDMWA [۱۰]

از دیگر کارهای انجام شده وب‌کاوی، در حوزه داده‌های مکانی می‌توان به استخراج Landmarks از داده‌های تحت وب اشاره کرد [۱۱] و [۱۲]. لندمارک‌ها (شکل ۳) اشیایی بارز و دائمی اند که نقش مهمی در راهنمایی مسافران ایفا می‌کنند و معمولاً مورد توجه توریست‌ها قرار دارند، به طور مثال می‌توان به برج میلاد و برج آزادی در شهر تهران اشاره داشت. وب‌کاوی ابزار جدیدی برای آشکارسازی لندمارک‌ها در اختیار ما قرار می‌دهد. Zheng و همکاران [۱۱] ذکر می‌کنند که تقریباً لیست کاملی و قابل اتکایی از لندمارک‌ها در سرتاسر جهان، وجود ندارد و نیز به دست آوردن مدل بصری (مانند یک مدل سه بعدی) قابل اعتماد برای هر لندمارک کاری مشکل است. آنها برای حل این مساله از مقدار بسیار بزرگ داده‌های چندرسانه‌ای در وب، موتورهای جستجوی تصاویر اینترنتی، توسعه ابزارهای تشخیص اشیاء و تکنیک‌های خوشه بندی استفاده کرده‌اند. ابتدا یک لیست جامع از لندمارک‌های احتمالی ایجاد شد. این لیست با کاوش دو منبع به دست آمد: ۱- حدود ۲۰ میلیون تصاویر ژئوتگ شده به وسیله GPS و ۲- صفحات آنلاین تورهای راهنمای مسافران. سپس با استفاده از ابزارهای داده‌کاوی لندمارک‌ها استخراج شدند (شکل ۴). در نتیجه این پروژه ۵۳۱۲ لندمارک از ۱۲۵۹ شهر در ۱۴۴ کشور به دست آمد.



شکل ۳- مثال‌هایی از لندمارک‌ها در جهان [۱۱]



شکل ۴- چارچوب کلی فرآیند استخراج لندمارک‌ها [۱۱]

Tezuka و همکاران [۱۳] تلفیق نزدیک‌تر موتورهای جستجو و سیستم اطلاعات جغرافیایی را مورد بررسی قرار داده‌اند. موتورهای جستجو بزرگ و پرتال‌ها^۱ مانند Ask، Yahoo، Google و MSN امکان جستجو بر حسب موقعیت را فراهم می‌کنند. در حالی که خدمات این سرویس‌ها در بسیاری از موارد برای کاربران مناسب است، ولی قابلیت‌های محدودی در موقعیت‌های دیگر دارد. به طور مثال کاربری که در حال رانندگی به دنبال یک رستوران می‌گردد، نمی‌تواند به نقشه توجه کند، او ترجیح می‌دهد، آدرس رستوران مورد نظر خود را بشنود. کاربر دیگری دوست دارد، لیستی از رستوران‌ها را در یک منطقه خاص به دست بیاورد. کاربر دیگر مایل است، نظرات مردم یک منطقه را با توجه به آنچه در وب منتشر می‌شود، به دست بیاورد. چنین اطلاعاتی به وسیله موتورهای جستجوی فعلی فراهم نمی‌شوند. چنین اطلاعاتی را می‌توان به وسیله تلفیق نزدیک‌تر موتورهای جستجو و GIS به دست آورد. مقاله تلفیق را در سه حوزه استخراج اطلاعات مکانی، کشف دانش و ارائه اطلاعات مکانی مورد بررسی قرار می‌دهد. Yang و Claramunt [۱۴] یک رویکرد ترکیبی برای شخصی‌سازی وب مکانی ارائه داده‌اند. هدف شخصی‌سازی کاهش جریان اطلاعات، تطبیق عرضه بر نیازها و علایق کاربر و در نتیجه بهبود رضایت او است. چارچوب پیشنهاد شده، اعمال کاربر بر روی موجودیت‌های مکانی و موقعیت‌های مرجع موجود بر روی وب‌سایت مدنظر را پیگیری می‌کند، و مناسب‌ترین موجودیت‌های مکانی را ارائه می‌دهد که با ترجیحات کاربر تطبیق پیدا می‌کنند. علایق و ترجیحات کاربر از این فرض به دست می‌آید که به یک موجودیت مکانی ارزش بالاتری تعلق می‌گیرد، اگر به موجودیت مکانی مورد علاقه کاربر شبیه هو ب آن نزدیک باشد. این روش شباهت معنایی، مجاورت مکانی و زنجیره مارکوف را ترکیب می‌کند، تا موجودیت مکانی بعدی را ارائه دهد.

شباهت معنایی بازتاب دهنده این است که یک موجودیت تا چه حدی به موجودیت دیگر در حوزه معنایی نزدیک است. مجاورت مکانی یک شکل مفهومی از معکوس فاصله میان دو موجودیت را تخصیص می‌دهد. زنجیره مارکوف به صورت ضمنی ردپای کاربر را در وب رهگیری و ثبت می‌کند، و الگوهای ناوبری و دانش را به منظور پیش‌بینی تعاملات کاربر در وب استنتاج می‌کند.

Collier و همکاران [۱۵] سیستم BioCaster را توسعه داده‌اند، که یک سیستم متن‌کاوی مبتنی بر آنتولوژی برای کشف و پیگیری توزیع شیوع بیماری‌های واگیر از سیگنال‌های زبان‌شناسی در وب است. سیستم به طور پیوسته اسناد گزارش شده از بیش از ۱۷۰۰ فید RSS را تحلیل می‌کند، آنها را برای موضوعات مرتبط دسته‌بندی می‌کند، سپس این موضوعات بر روی نقشه Google با استفاده از اطلاعات ژئوکد شده رسم می‌شوند.

^۱ - وب‌سایتی که به عنوان یک بزرگراه ورودی به اینترنت عمل میکند. مانند Yahoo و Excite و HotBot

سیستم متشکل از ۴ مرحله اصلی است: دسته‌بندی موضوعات، شناسایی موجودیت‌های نام‌دار، کشف بیماری/موقعیت و شناسایی رویداد. تحلیل رویدادهای مراتب بالاتر برای تشخیص دقیق‌تر سیگنال‌های هشدار استفاده می‌شود. به صورت خلاصه می‌توان سیستم را این گونه توصیف کرد: منابع از فید خبرخوان جمع‌آوری می‌شود ← پیش پردازش می‌شود ← خوشه‌بندی می‌شود ← دسته‌بندی خودکار صورت می‌گیرد ← ۱۸ نوع موجودیت بر اساس آن‌تولوژی BioCaster شناسایی می‌شود ← جفت موقعیت-بیماری بر روی یک پرتال عمومی پلات می‌گردد.

در استخراج رویدادها، سطوح اضافی از تحلیل معنایی به سیستم اجازه می‌دهد به یک سطح عمیق‌تر از فهم درباره میزان اهمیت رویداد برای سلامت عمومی دست یابد. چنین فهمی از طریق قوانین روشن کردن زمینه‌های مکانی-زمانی و زبان‌شناسی حاصل می‌شود. این قوانین قابلیت تطبیق کلاس‌ها، رشته حروف، عبارات متداول و نوع موجودیت را دارد.

بعضی از محققان با استفاده از داده‌کاوی یک کتابخانه از نقشه‌های قابل جستجو ایجاد کرده‌اند. این نقشه‌ها از درون مقالات استخراج شده‌اند. چنین نقشه‌هایی به سختی قابل دسترسی‌اند، چرا که به ندرت جدا از خود مقالات فهرست شده‌اند. سپس متادیتای مرتبط با این نقشه‌ها با استفاده از زیرنویس نقشه، کلمات موجود در نقشه، عنوان مقاله و جملات داخل مقاله که به نقشه ارجاع می‌دهند، استخراج گردید. سپس نقشه‌ها بر اساس موضوع، زمان و منطقه دسته‌بندی شدند. نمونه اولیه از پیاده‌سازی این سیستم را می‌توان در وب یافت^۱ [۱۶].

Miron و همکاران [۱۷] یک هستی‌شناسی مکانی پیشنهاد داده‌اند، که می‌تواند برای الصاق خصوصیات مکانی به منابع وب استفاده شود، این هستی‌شناسی همراه با یک استنتاج‌گر مکانی ارائه شده است. چنین کاری در حوزه تحلیل‌های معنایی قرار می‌گیرد، که یک مدل جدید بر ای جستجو در وب معنایی است. که به سولاتی نظیر "آیا نمونه x با نمونه y ارتباطی دارد؟" در این تحقیق همچنین راه‌های مختلف استخراج اطلاعات مکانی در فرآیند اکتشاف وابستگی‌های معنایی بررسی شده است. در نهایت چگونگی استفاده از مفاهیم مکانی برای محدود کردن فضای جستجو به منطقه مکانی داده شده مورد ارزیابی قرار گرفته است.

Huang [۱۸] از وب‌کاوی و هستی‌شناسی مکانی برای بهبود کیفیت سرویس WebMap استفاده کرده است. بهبود کیفیت سیستم پرس‌وجو از دو جنبه صورت گرفته است: کشف و استخراج اطلاعات آدرس‌ها از وب به کمک هستی‌شناسی به منظور گسترش پایگاه داده ژئوکد شده، و بهبود قابلیت پرس‌وجو به وسیله تحلیل و استدلال مکانی بر روی پایگاه‌داده توسعه یافته.

Klien و Grcar [۱۹] یک روش برای خودکارسازی برچسب‌گذاری معنایی سرویس‌های مکانی ارائه داده‌اند. این محققان روش خود را مبتنی بر تکنیک‌های متن‌کاوی توسعه داده‌اند.

رشد مقدار داده‌های نیمه‌ساختاریافته و ساختارنیافته در سایت‌های توریسم با طراحی ناهمگن، نیازمند مکانیسم‌های استخراج اطلاعات به طور مثال برای ایجاد پرتال‌های توریسم است. به منظور ساخت محیط معنایی eTourism جمع‌آوری قیمت اتاق‌ها یک نیاز خاص است. قیمت اتاق‌ها و اطلاعات مرتبط با آن اغلب در یک ساختار جدولی ظاهر می‌شوند، که هنوز یک چالش برای استخراج اطلاعات از وب است. Buttinger و همکاران [۲۰] در

¹ - <http://scilsresx.rutgers.edu/~gelern/maps/>

مطالعه خود برای برخورد با این چالش، ابتدا الگوهای جدول قیمت‌های مختلف را شناسایی کرده‌اند، که به وسیله موقعیت یک تعداد از خصوصيات توصیف شده‌اند، که قیمت اتاق را مشخص می‌کنند. سپس یک مدل هستی‌شناسی برای توصیف قیمت‌های توريسم توسعه داده شده است.

تعداد زیادی از کاربران در به اشتراک‌گذاری اطلاعات در وب از طریق خدمات سایت‌های میزبان محتوا و شبکه‌های اجتماعی شرکت می‌کنند. این کاربران محتوایی تولید می‌کنند، که به وسیله عبارت‌هایی برچسب‌گذاری شده است که محتوا را از نگاه کاربر توصیف می‌کنند. مجموعه‌های بزرگ از تصاویر برچسب‌گذاری شده در سایت‌های میزبان تصویر، از نظر توصیف‌های معنایی و شاخص‌های مکانی غنی‌اند. علاوه بر این ژئوتگ‌ها که داده‌های موقعیتی تولید شده توسط ماشین‌اند، اغلب همراه با این تصاویر ارائه شده‌اند. Deng و همکاران [۲۱] روشی توسعه داده‌اند، که از تحلیل تگ‌ها و ژئوتگ‌ها برای شناسایی و توصیف موقعیت‌های ضمنی موجود در مجموعه تصاویر کاربران استفاده می‌کند. این روش همچنین اجازه شناسایی معنای ارائه شده توسط کاربران درباره مکان‌های موجود در مجموعه تصاویرشان را می‌دهد. ابتدا تصاویر ژئوتگ شده خوشه‌بندی می‌شوند، سپس یک تحلیل هم‌زمانی بر روی تگ‌های داخل هر خوشه صورت می‌گیرد تا مفهوم‌سازی مکانی موجود در داخل خوشه استخراج شود. مفاهیم استخراج شده لزوما دارای طبیعت مکانی نیستند (مانند نام هواپیماها یا خطوط هوایی در تصاویر اخذ شده در محوطه یک فرودگاه) بنابراین در مقایسه با مفاهیم استخراج شده به وسیله استفاده ساده از منابع موقعیتی موجود (مانند اطلس‌ها) دارای مزایایی است.

Sheng و همکاران [۲۲] روشی برای استخراج علایق خاص مکانی از داده‌های کلیک وب توسعه داده‌اند، که می‌تواند تغییرات حاصل در منطقه مورد مطالعه را نیز کشف کند. در حوزه وب‌کاوی مکانی، می‌توان به این موارد نیز اشاره کرد: استخراج روابط توپولوژیک از وب [۲۳]، کاوش لاگ‌های وب سایت برای بهبود زمان پاسخ سیستم WebGIS [۲۴]، استخراج خصوصیات شهری با استفاده از وب‌کاوی [۲۵] و [۲۶]، جمع‌آوری داده‌های مکانی-زمانی به طور مثال روند گسترش آتش‌سوزی با استفاده از شبکه‌های اجتماعی مبتنی بر موقعیت^۱ [۲۷]، استخراج رویدادهای مکانی نظیر محل نبردهای جنگ جهانی دوم با استفاده از وب‌کاوی [۲۸]، و استخراج نام مکان‌ها (شامل نام‌های بومی و غیر رسمی مانند مرکز شهر، بالاشهر و...) به منظور بهبود اطلس‌های جغرافیایی [۲۹] و [۳۰]. در ادامه مورد اخیر یعنی استخراج نام مکان‌ها را بیشتر بررسی خواهیم کرد.

استخراج نام مکان‌ها از اسناد ویکی‌پدیا

نام مکان‌ها نقشی کلیدی در فرمول‌بندی پرسش و پاسخ‌ها برای بازیابی اطلاعات جغرافیایی بازی می‌کنند، این نقش وقتی پای وب به میان می‌آید بیشتر به چشم می‌خورد. با هدف بهبود کیفیت اطلس‌ها برای سیستم‌های بازیابی اطلاعات جغرافیایی، یک روش برای شناسایی نام مکان‌ها پیشنهاد می‌شود که از اطلاعاتی استفاده می‌کند که مردم در وب به اشتراک می‌گذارند. ما بررسی کردیم که مردم چگونه با استفاده از عبارات مکانی به یک مکان در پرسش و پاسخ‌های وب ارجاع می‌دهند، و از وب‌کاوی به عنوان وسیله‌ای برای تصمیم‌گیری درباره اینکه آیا موجودیت نامگذاری شده در واقعیت یک مکان است یا خیر استفاده کردیم.

¹ - Location Based Social Network e.g. <http://www.foursquare.com>

شناسایی نام مکان‌های جدید یک جنبه حیاتی در ساخت و حفظ سرویس‌های اطلس جغرافیایی است. ما داده‌های کاوش شده از وب را برای طبقه‌بندی نام‌های استخراج شده به عنوان نام مکان‌ها استفاده کردیم. با کاوش وب سایت ویکی‌پدیا^۱ یک لیست اولیه از نام‌ها ایجاد شد. برای نام شهرها یک پرسش‌وپاسخ با استفاده از عبارات مکانی ایجاد شد، و اسنادی که بازگردانده شدند، ذخیره شدند. از تکنیک‌های وب‌کاوی برای کاوش این اسناد به منظور یافتن نام‌های دیگر یک شهر و یا نام‌های قدیمی آن استفاده شد.

هدف این پژوهش استخراج نام مکان‌ها به منظور تست تکنیک وب‌کاوی است. می‌خواهیم نام‌های دیگر و یا قدیمی شهرهای ایران را با استفاده از اطلاعات ویکی‌پدیا استخراج کنیم. به همین منظور ابتدا فهرستی صد تایی از نام شهرهای ایران تهیه شد، سپس یک پرسش و پاسخ به صورت زیر ایجاد شد که سند مربوط به نام شهر را در صورت وجود در ویکی‌پدیا می‌یابد، این صفحه به منظور تحلیل‌های بعدی ذخیره می‌گردد.

[http://fa.wikipedia.org/wiki/\[cityName\]](http://fa.wikipedia.org/wiki/[cityName])

صفحات ویکی‌پدیا فارسی در مورد شهرهای ایران دارای ساختار منظم و استاندارد نیست. گاهی نام پیشین شهر در بخشی به نام "پیشینه" آمده است، گاهی این نام‌ها در قسمت "تاریخچه" ذکر شده و گاهی در بخش‌های "نام (نام‌های پیشین)"، "نام‌گذاری"، یا "وجه تسمیه" طبقه‌بندی شده‌اند (شکل ۵)، این امر روند استخراج نام‌ها را دشوار می‌کند. برای رفع این مشکل در یک صفحه هر کدام از این پنج بخش جستجو می‌شود و در صورتی که یک یا تعداد بیشتری از این بخش‌ها را بیابد، متن آن بخش‌ها از صفحه جدا و در یک فایل متنی به نام آن شهر ذخیره می‌شود. اگر این بخش‌ها یافت نشود صفحه برای تحلیل‌های بیشتر به حال خود رها می‌گردد.

در منطقه، حاکم بودن شرایط مناسب اوضاع توپوگرافی، وجود اراضی مسطح با شیب ۲٪ در منطقه کوهستانی و بالاخره انوم مرکز همدانی تولیدات کشاورزی و ارائه خدمات متقابل به روستاهای حوزه نفوذ به ویژه ایالات پنجگانه مستقر در منطقه بوده‌است می‌رساند که این پدیده در شکل‌گیری شهر و محله بندی آن عامل تعیین کننده می‌باشد.

محتویات [بهبود]

- ۱ پیشینه
- ۲ آثار تاریخی
- ۳ محله‌های اصلی
- ۴ آموزش
- ۴٫۱ مدارس
- ۵ مراکز آموزشی
- ۶ نقاط دیدنی
- ۷ صنایع دستی
- ۸ سوغات
- ۹ مشاهیر زنجان
- ۱۰ روزنامه‌ها
- ۱۱ هفته‌نامه‌های زنجان
- ۱۲ هتل‌ها
- ۱۳ دیدنی‌های زنجان
- ۱۴ شهرهای خواهرخوانده
- ۱۵ جستارهای وابسته
- ۱۶ پیوند به بیرون
- ۱۷ منابع

پیشینه [ویرایش]

در مورد وجه تسمیه شهر زنجان تاکنون نظرات گوناگونی از سوی پژوهشگران و نویسندگان ارائه شده است. قدیمی‌ترین نامی که به این منطقه اطلاق شده است زندیگان به معنای اهل کتاب زند (معروف‌ترین کتاب آیین زردشتی ساسانی) و گان از پساوند فارسی عهد باستان است که در دوره ساسانیان بر این منطقه گذاشته شده است. گفته می‌شود بنای شهر زنجان در زمان اردشیر بابکان ساخته شده و در آن زمان نام شهرین یعنی منسوب به شاه به آن اطلاق می‌شده است. از اواخر دوره قاجاریه به علت استقرار ایل **خمسه** فارس نام **خمسه** نیز بر آن نهاده شد.

شکل ۵- نمونه‌ای از تقسیم‌بندی صفحات ویکی‌پدیا برای شهرهای ایران

حال با استفاده از تکنیک‌های متن کاوی فایل‌های متنی ذخیره شده را مورد کاوش قرار می‌دهیم. ابتدا کلمات انگلیسی را حذف می‌کنیم. سپس کلمات بولد و یا ایتالیک و یا قرار گرفته در بین علامت‌های « » را جدا می‌کنیم،

¹ - <http://fa.wikipedia.org>

معمولا چنین کلماتی نشان‌دهنده نام شهرها هستند. سپس با این فرض پیش می‌رویم که با حذف کلمات و حروف اضافه به کمک پایگاه‌داده‌های کلمات فارسی (مانند آنچه که شورای عالی اطلاع‌رسانی تهیه کرده است)^۱ آنچه باقی خواهد ماند، اسامی معرفه خواهند بود. که شامل نام شهرها، نام نویسندگان، کتاب‌ها و نام پیشینیان است. در این مرحله نام نویسندگان، کتاب‌ها و پیشینیان (مانند یاقوت حموی، اردشیر ساسانیان) با استفاده از فهرست اعلام حذف می‌شوند، و نام شهرها باقی می‌ماند. باید نام فعلی شهرها را حذف کنیم چرا که گاهی عباراتی مانند "کرج گاهی از روستاهای طالقان یا شهرستانک محسوب می‌شده است" در متن وجود دارد. اگر چنین پردازشی انجام نشود نام‌های طالقان و شهرستانک جزو نام‌های دیگر کرج محسوب خواهد شد که اشتباهی فاحش است. به همین دلیل با استفاده از پایگاه نام‌های جغرافیایی ایران^۲ نام فعلی شهرها، رودخانه‌ها، کوه‌ها و ... تا آنجا که ممکن است از متن حذف می‌شود. آنچه باقی خواهد ماند نام قدیمی شهرها خواهد بود. در جدول ۱ نمونه‌ای از نتایج نشان داده شده است:

همان‌طور که در نام‌های دیگر شهر اصفهان مشاهده می‌شود، این مجموعه نام‌ها ممکن است شامل نام‌های دیگری نیز باشند که نام شهر محسوب نمی‌شوند ولی در این مجموعه قرار گرفته‌اند (False negative error). همچنین احتمال دارد بعضی از نام‌های شهر در متن وجود داشته باشد که طی فرآیند پردازش متن حذف شده باشند (False positive error). برای کاهش چنین خطاهایی لازم است که پردازش‌های پیشرفته‌تری بر روی متن انجام شود و از پایگاه‌داده‌های بهتری استفاده گردد.

جدول ۱ - نمونه ای از اطلاعات استخراج شده

نام شهر	نام‌های دیگر یا پیشین
زنجان	زندگان، شهین، خمسه، زنگان
آشتیان	ابرشتگان، ابرشتجان، ورّه
آستارا	آسترو، هُسترو، آهسترو
آبادان	عَبّادان، آیفانا، آیفادانا، آیاتان
کرج	کراچ، کرژ
اصفهان	گابای، گی، جی، اسپهان، اسپاهان، گابای، تابای، یهودیه، واسپوهران
بابل	بارفروش، باوُل
بندرعباس	سورو، گمبرون، سارو، خامه ور رون، خمبرو، سالمونت، تیاب، جرون
تهران	تهرام، طهران
چالوس	سالوس، شالوس، چاله‌است، چاله‌اس، چارلس
خرمشهر	خاراگس، بیان، کوت المَحْمَرّه، المحمره، خونین‌شهر
دزفول	آوان، دزپِل، دژپِل، دزفیل، دسفیل، اندامش
رشت	رشت، بیه، دارالمرز
زابل	سیستان، نیمروز، اریاپولیس
سبزوار	بیّهق، خسروگرد

^۱ - http://www.prosody.ir/index.php?option=com_content&view=article&id=46&Itemid=18

^۲ - <http://geonames.ncc.org.ir/HomePage.aspx?TabID=4572&Site=geonames.ncc.org&Lang=fa-IR>

نتیجه‌گیری و پیشنهادات

وب‌کاوی حوزه تحقیقاتی بسیار فعالی است. یک تحقیق مانند آنچه در این نوشتار ذکر شده است، تنها بخش کوچکی از این محدوده را پوشش می‌دهد. در واقع تلاش این نوشته ایجاد یک نقطه شروع برای مطالعات بیشتر در این حوزه و به خصوص ارتباط آن با استخراج اطلاعات مکانی و بهبود وب‌سرویس‌های مکانی است. در این پژوهش مشاهده شد که اطلاعات وب که اغلب به صورت داوطلبانه فراهم شده‌اند، می‌تواند منبع بزرگی برای استخراج اطلاعات مکانی باشد. در این پژوهش با دقت خوبی نام‌های قدیمی عده‌ای از شهرهای ایران استخراج شد. برای رسیدن به دقت‌های بهتر باید از پایگاه‌داده‌های روزرسانی شده و همچنین تکنیک‌های پیشرفته‌تر متن‌کاوی استفاده شود.

در حوزه مکانی، اغلب روش‌ها بر روی استخراج داده از اطلاعات مکانی متمرکز شده‌اند. این روش‌ها تا حدی وابسته به زبان اند. پس این ضرورت وجود دارد که چنین سیستم‌هایی مختص زبان فارسی توسعه یابند، تنها BioCaster با استفاده از مترجم‌های آنلاین سعی در استخراج اطلاعات از منابع فارسی داشت، ولی نتایج حاصل از آن به دلیل ضعف سیستم‌های ترجمه و در نظر نگرفتن هستی‌شناسی زبان فارسی دقت پایینی دارد. در این زمینه می‌توان استخراج نام‌های غیر رسمی مکان‌ها در زبان فارسی را نیز در نظر گرفت. با وجود حجم بزرگ داده‌های مکانی در قالب‌های برداری و رستری که در وب موجود می‌باشند، به ندرت روشی برای استخراج اطلاعات مکانی از این منابع توسعه داده شده است، استخراج خودکار اطلاعات مکانی از این منابع می‌تواند مورد مطالعه بیشتری قرار بگیرد.

از آنجا که امروزه بسیاری از رویدادها در وب بارتاب می‌یابد، می‌توان از وب‌کاوی برای به روز رسانی نقشه‌ها استفاده کرد، به طور مثال افتتاح پروژه‌ها، مراکز تجاری، اماکن فرهنگی و غیره بلافاصله در اخبار بازتاب می‌یابد، و امکان ارزشمندی برای به روزرسانی سریع نقشه‌ها با استفاده از وب‌کاوی فراهم می‌شود. از طرف دیگر می‌توان از اخبار جرایم بازتاب یافته در مطبوعات و رسانه‌های دیگر برای بررسی مکانی جرم استفاده کرد. چنین سیستم‌هایی باید برای زبان فارسی توسعه یابند.

مراجع:

- [1] Liu, B., & Chen-Chuan-Chang, K. (2004). Special Issue on Web Content Mining. *ACM SIGKDD Explorations Newsletter*, 6 (2).
- [2] Jones, C. B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M., et al. (2002). Spatial information retrieval and geographical ontologies an overview of the SPIRIT project. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. Tampere, Finland: ACM.
- [3] Borges, K. A., Laender, A. H., Medeiros, C. B., & Davis, C. (2007). Discovering geographic locations in web pages using urban addresses. *Proceedings of the 4th ACM workshop on Geographical information retrieval*. Lisbon, Portugal: ACM.
- [4] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web Usage Mining Discovery and Applications of Usage. *ACM SIGKDD Explorations Newsletter*, 1 (2), 12-23.

- [5] Kosala, R., & Blockeel, H. (2000). Web mining research: a survey. *ACM SIGKDD Explorations Newsletter* , 2 (1), 1-15.
- [6] Kolari, P., & Joshi, A. (2004). Web Mining: Research and Practice. *Computing in Science & Engineering* , 6 (4), 49-53.
- [7] Srivastava J, D. P. (2005). Web Mining Concepts Applications and Research Directions. In W. Chu, & T. Young Lin, *Foundations and Advances in Data Mining-Studies in Fuzziness and Soft Computing* (Vol. 180/2005, pp. 275-307). Heidelberg, Berlin: Springer.
- [8] Berendt, B., Hotho, A., Mladenic, D., van Someren, M., Spiliopoulou, M., & Stumme, G. (2004). A Roadmap for Web Mining: From Web to Semantic Web. In *Lecture Notes in Computer Science* (Vol. 3209/2004, pp. 1-22). Heideberg, Berlin: Springerlink.
- [9] Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal.* , 69 (4), 211-221.
- [10] Pultar, E., Raubal, M., & Goodchild, M. (2008). Geospatial Exploratory Data Mining Web Agent. *16th ACM SIG SPATIAL international conference on Advances in geographic information systems*. Irvine, California: ACM.
- [11] Zheng, Y.-T., Zhao, M., & Song, Y. (2009). Tour the world: Building a web-scale landmark recognition engine. *Computer Vision and Pattern Recognition 2009* (pp. 1085 - 1092). Miami, FL: IEEE.
- [12] Tezuka, T., & Tanaka, k. (2005). Landmark Extraction: A Web Mining Approach. In A. G. Cohn, & D. M. Mark, *Spatial Information Theory-Lecture Notes in Computer Science* (Vol. 3693/2005, pp. 379-396). Heidelberg, Berlin: Springer.
- [13] Tezuka, T., Kurashima, T., & Tanaka, k. (2006). Toward Tighter Integration of Web Search with a Geographic Information System. *15th international conference on World Wide Web* (pp. 277-286). Edinburgh, Scotland: ACM.
- [14] Yang, Y., & Claramunt, C. (2005). A Hybrid Approach for Spatial Web Personalization. In K.-J. Li, & C. Vangenot, *Web and Wireless Geographical Information Systems-Lecture Notes in Computer Science* (Vol. 3833/2005). Berlin,Heidelberg: Springer.
- [15] Collier, N., Doan, S., & Kawazoe, A. (2008). BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics* , 24 (24), 2940-2941.
- [16] Gelernter, J. (2009). Data mining of maps and their automatic region-time-theme classification. *1* (1), 39-44.
- [17] Miron, A., Gensel, J., & Villanova-Oliver, M. (2009). Discovery of spatial links between individuals on the Geospatial Semantic Web. *Geoinformatics, 2009 17th International Conference* (pp. 1-6). Fairfax, VA: IEEE.
- [18] Huang, M. (2010). Use geographic ontology and web mining to enhance service quality of web map. *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference* (pp. 625 - 629). Chengdu: IEEE.
- [19] Grcar, M., & Klien, E. (2007). Using Term-matching Algorithms for the Annotation of Geo-services. *Web Mining 2.0, Workshop at ECML*. Warsaw, Poland.

- [20] Buttinger, C., Feilmayr, C., Guttentbrunner, M., Parzer, S., & Pröll, B. (2010). Extracting Room Prices from Web Tables — an Ontology-Aware Approach. *Information and Communication Technologies in Tourism 2010*, 6, 223-234.
- [21] Deng, D.-P., Chuang, T.-R., & Lemmens, R. (2009). Conceptualization of place via spatial clustering and co-occurrence analysis. *Proceedings of the 2009 International Workshop on Location Based Social Networks*. ACM.
- [22] Sheng, C., Hsu, W., & Li Lee, M. (2008). Discovering geographical-specific interests from web click data. *Proceedings of the first international workshop on Location and the web* (pp. 41-48). Beijing, China: ACM.
- [23] Schockaert, S., Smart, P., Abdelmoty, A., & Jones, C. (2008). Mining Topological Relations from the Web. *International Workshop on Database and Expert Systems Application* (pp. 652-656). Turin, Italy: IEEE.
- [24] Holowczak, R., Taksa, I., & Levine, I. (2007). Mining Web Logs to Improve WebGIS Application Response Times. *International Conference on Information Technology*. Las Vegas, Nevada, USA: IEEE.
- [25] Fujisaka, T., Lee, R., & Sumiya, K. (2010). Exploring urban characteristics using movement history of mass mobile micro bloggers. *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications* (pp. 13-18). Annapolis, Maryland: ACM.
- [26] Li, M., & Chen, A. (2008). A Web Mining Based Measurement and Monitoring Model of Urban Mass Panic in Emergency Management. *Fifth International Conference on Fuzzy Systems and Knowledge Discovery* (pp. 366 - 370). Jinan Shandong: IEEE.
- [27] Longueville, B., Smith, R., & Luraschi, G. (2009). "OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. *International Workshop on Location Based Social Networks* (pp. 73-80). Seattle, Washington: ACM.
- [28] Xu, k., Li, R., Bao, S., Han, D., & Yu, Y. (2008). SEM: Mining Spatial Events from the Web. In T. Washio, E. Suzuki, K. Ting, & A. Inokuchi, *Advances in Knowledge Discovery and Data Mining-Lecture Notes in Computer Science* (Vol. 5012/2008, pp. 393-404). Berlin, Heidelberg: Springer-Verlag.
- [29] Jones, C., Purves, R., Clough, P., & Joho, H. (2008). Modelling vague places with knowledge from the Web. 22 (10), 1045-1065.
- [30] Twaroch, F., Smart, P., & Jones, C. (2008). Mining the Web to Detect Place Names. *2nd international workshop on Geographic information retrieval*. Napa Valley, California, USA: ACM.