دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

خوشه بندی کاربران وب و واکشی اولیه صفحات وب

با استفاده از آنالیز معنایی پنهان احتمالاتی

پایان نامه کارشناسی ارشد مهندسی کامپیوتر- نرم افزار

فاطمه زهرا حیدری

استاد راهنما

دکتر محمدعلی منتظری

دکتر احمد براآنی

۱۳۹۳

دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

پایان نامه کارشناسی ارشد مهندسی کامپیوتر- نرم افزار خانم فاطمه زهرا حیدری

تحت عنوان

خوشه بندی کاربران وب و واکشی اولیه صفحات وب
با استفاده از آنالیز معنایی پنهان احتمالاتی

در تاریخ ۱۳۹۳/۰۶/۲۶ توسط کمیته تخصصی زیر مورد بررسی و تصویب نهایی قرار گرفت

۱- استاد راهنمای پایان نامه	دکتر محمدعلی منتظری
۲- استاد راهنمای پایان نامه	دکتر احمد براآنی
۳- استاد داور	دکتر عبدالرضا میرزایی
۴- سرپرست تحصیلات تکمیلی دانشکده	دکتر محمدعلی خسروی‌فرد

تقدیم به
پدرم به استواری کوه
مادرم به زلالی چشمه
همسرم به صمیمیت باران

فهرست مطالب

عنوان **صفحه**

فصل اول : مقدمه

۱–۱ انگیزه و پیش زمینه ها ۱

۱–۲ هدف پژوهش ۳

۱–۳ ساختار پژوهش ۳

فصل دوم: استخراج اطلاعات از داده های کاربرد وب

۲–۱ مقدمه ۵

۲–۲ داده کاوی و استخراج دانش ۵

۲–۳ وب کاوی ۷

۲–۴ چالش های وب کاوی ۸

۲–۵ اجزای اصلی وب کاوی ۸

۲–۵–۱ محتوا کاوی وب ۹

۲–۵–۲ ساختارکاوی وب ۹

۲–۵–۳ کاربرد کاوی وب ۹

۲–۶ آماده سازی و مدل سازی داده ۱۱

۲–۶–۱ منابع داده ۱۱

۲–۶–۲ آماده سازی و پیش پردازش داده ها ۱۴

۲–۷ کشف الگو از داده های کاربرد وب ۲۰

۲–۸ سطوح و انواع تحلیل ۲۲

۲–۹ کاربرد های کاربرد کاوی وب ۲۳

۲–۱۰ جمع بندی ۲۴

فصل سوم : خوشه بندی کاربران وب

۳–۱ مقدمه ۲۵

۳–۲ خوشه بندی ۲۵

۳–۳ مروری بر کارهای انجام شده در زمینه خوشه بندی کاربران وب ۲۸

۳–۴ روش های براساس عامل های پنهان ۳۱

۳-۴-۱ تکنیک تحلیل معنایی پنهان ۳۲
۳-۴-۲ تکنیک نمایه سازی تصادفی ۳۵
۳-۴-۳ تکنیک تحلیل معنایی پنهان احتمالاتی ۳۷
۳-۴-۴ کشف و آنالیز الگوهای کاربردی با استفاده از تحلیل معنایی پنهان احتمالاتی ۴۱
۳-۵ مروری بر کارهای انجام شده در زمینه خوشه بندی کاربران وب براساس عامل های پنهان ۴۶
۳-۶ برآورد تعداد عامل های پنهان ۴۸
۳-۷ ارزیابی روش های خوشه بندی ۴۹
۳-۸ جمع بندی ۵۲
فصل چهارم: روش پیشنهادی برای خوشه بندی کاربران وب براساس تحلیل معنایی پنهان احتمالاتی
۴-۱ مقدمه ۵۳
۴-۲ خوشه بندی کاربران وب براساس آنالیز احتمالاتی معنایی پنهان ۵۴
۴-۲-۱ پیش پردازش داده ۵۴
۴-۲-۲ مدل سازی کاربر براساس تحلیل معنایی پنهان احتمالاتی ۵۵
مجموعه گردشگری کاربران شخصی ۵۵
تقسیم بندی URLها ۵۶
۴-۳ اعمال تحلیل معنایی پنهان احتمالاتی با استفاده از توابع وزن دهی مختلف برای هر کاربر ۵۷
۴-۴ خوشه بندی الگوهای کاربری ایجاد شده ۶۰
۴-۵ ایجاد پروفایل عمومی کاربران ۶۲
۴-۶ معرفی روش ها به منظور مقایسه ۶۳
۴-۷ جمع بندی ۶۳
فصل پنجم: ارزیابی روش پیشنهادی و کاربرد آن
۵-۱ مقدمه ۶۵
۵-۲ ارزیابی نتایج ۶۵
۵-۲-۱ پیش پردازش و منبع داده ۶۵
۵-۲-۱ بررسی تعداد عامل های پنهان با استفاده از توابع وزنی مختلف ۶۶
۵-۲-۲ اندازه گیری صحت خوشه بندی ۶۸
۵-۲-۳ ایجاد پروفایل های عمومی کاربر ۷۲

۵–۳ کاربرد : واکشی اولیه صفحات ۷۳
۵–۳–۱ قوانین واکشی اولیه ۷۴
۵–۳–۲ نتایج ۷۵
۵–۴ جمع بندی ۷۶
فصل ششم: نتیجه گیری
۶–۱ مقدمه ۸۱
۶–۲ ارزیابی الگوریتم پیشنهادی ۸۱
۶–۳ پیشنهاداتی برای کارهای آینده ۸۲
۶–۴ جمع بندی ۸۲
مراجع ۸۳

چکیده

در سالهای اخیر با توجه به رشد سریع وب جهانی، تحقیقات وسیعی پیرامون مدل کردن رفتار پیمایشیِ کاربران در وب‌سایت‌ها انجام گرفته است. در این راستا کاربردکاوی وب با هدف به‌دست آوردن الگوهای رفتار پیمایشی کاربران وب، مورد استفاده بسیاری از محققان قرار گرفته است.

به‌طورکلی کاربران وب رفتارهای متفاوتی متناسب با نیازهای اطلاعاتی و وظایف مورد علاقه خود در بازدید از وب‌سایت‌ها از خود نشان می‌دهند، تمامی رفتارهای پیمایشی کاربران در فایل‌های ثبت وب قابل ردیابی است. یکی از تکنیک‌های مورد استفاده در کاربرد کاوی وب خوشه‌بندی کاربران وب می‌باشد. در تکنیک خوشه‌بندی کاربرانی که رفتار پیمایشی مشابهی دارند در یک خوشه قرار می گیرند. هر خوشه منجر به ایجاد پروفایل‌های کاربری می‌شود که در برنامه‌های کاربردی مانند واکشی اولیه و حافظه نهان مورد استفاده قرار می‌گیرد. تکنیک‌های متداول و استاندارد کاربرد کاوی وب برای خوشه‌بندی کاربران‌وب می‌تواند الگوهای کاربردی را مستقیماً کشف کند، اما این تکنیک‌ها به‌طور خودکار نمی‌توانند مشخصات یا کیفیت عامل‌های پنهانی که منجر به کشف الگوهای پیمایشی مشترک می‌شوند را تعیین کنند. بنابراین نیاز به گسترش تکنیک‌هایی می‌باشد تا بتوان بصورت خودکار اهداف اساسی پیمایشی کاربران را شناسایی و رابطه معنایی پنهان میان کاربران‌وب و همچنین رابطه معنایی پنهان بین کاربران‌وب و اشیاء وب را استخراج کرد.

در این پژوهش، روشی براساس آنالیز معنایی پنهان احتمالاتی پیشنهاد می‌شود که مشخصات ذاتی رفتار پیمایشی کاربران را مشخص می‌کند. روش پیشنهادی فاکتورهای پنهان به‌دست‌آمده را جهت خوشه‌بندی الگوهای پیمایشی کاربران مورد استفاده قرار می‌دهد و پروفایل‌های کاربری را ایجاد می‌کند. نتایج خوشه‌بندی برای پیش‌بینی و واکشی درخواست‌های وب گروه‌های کاربران مورد استفاده و ارزیابی قرار می‌گیرد.

کارایی و برتری روش خوشه‌بندی کاربران از طریق آزمایش بر روی فایل‌های ثبت واقعی نشان داده می‌شود. روش پیشنهادی برای گروه‌بندی کاربران وب و واکشی اولیه با کارهای قبلی مقایسه شده و نتایج به‌دست آمده کارایی بهتر و میزان دقت بالای روش پیشنهادی را در مقایسه با روش‌های دیگر از خود نشان می‌دهد.

کلمات کلیدی: کاربرد کاوی وب ، خوشه‌بندی کاربران وب ، آنالیز معنایی پنهان احتمالاتی، واکشی اولیه صفحات

فصل اول

مقدمه

۱-۱ انگیزه و پیش‌زمینه‌ها

در دنیای امروز، وب رسانه‌ای محاوره‌ای و محبوب برای انتشار اطلاعات می‌باشد. درنتیجه، وب به بزرگترین بازار بدون مرز برای ارائه خدمات و اطلاعات و یا خرید و فروش کالا تبدیل شده است. وجود انبوه عرضه‌کنندگان سبب شده است که شرکت‌ها بر سر یک یک کاربرانی (مشتریانی) که از وب‌سایت آنها دیدار می‌کنند و شاید در مواقعی کالا و خدمات آنها را خریداری می‌کنند، با یکدیگر رقابت کنند. راز پیروزی در چنین رقابت سرسختانه‌ای اطلاع از نیازهای کاربران وب‌سایت و ارائه خدماتی که بتواند نیازها را برآورده سازد. در بیشتر موارد طراحی یک سایت براساس درک و تصور طراح از الگوهای کاربردی کاربران از وب‌سایت صورت می‌پذیرد. بنابراین اطلاع از الگوهای کاربردی کاربران منجر به انطباق هر چه بهتر خدمات ارائه شده توسط وب‌سایت مطابق با نیازهای کاربران و به تبع آن افزایش احتمال جذب کاربر می‌گردد. از سوی دیگر کاربران وب به دلیل رشد سریع اطلاعات و تعداد زیاد کاربران با مشکل سربار اطلاعات مواجه هستند. در نتیجه اینکه چگونه اطلاعات مورد نیاز کاربران در اختیار آنها قرار داده شود نیز به موضوعی بحرانی در بازیابی اطلاعات و مدیریت داده‌ها براساس وب تبدیل شده است.

راه‌حل‌های فراوانی براساس راه‌حل‌های تکنیکی برای حل مشکلات بیان شده ارائه شده است. در میان راه‌حل‌های ارائه شده، وب کاوی[1] به عنوان راه‌حلی کارآمد برای کشف ساختار محتوای وب، رفتار گردشگری کاربران براساس علاقه و تعاملات اساسی بین کاربران وب و اشیاء وب مورد استفاده قرار می‌گیرد.

وب‌کاوی فرایند کشف ارتباط ذاتی میان داده‌های وب می‌باشد، داده‌های وب به صورت اطلاعات متنی، پیوندی[2] و کاربردی است که از طریق آنالیز ویژگی‌های داده‌های وب و با استفاده از تکنیک‌های داده‌کاوی الگوهای جالبی را می‌توان کشف و استخراج کرد. با توجه به داده‌های مورد استفاده، وب‌کاوی شامل محتوا کاوی وب ، ساختار کاوی وب و کاربرد کاوی وب می‌باشد. در این پروژه، تمرکز اصلی بر روی کاربرد کاوی وب می‌باشد. همانطور که در فصل دو توضیح داده می‌شود، کاربرد کاوی وب بکارگیری روش‌های داده‌کاوی بر روی داده‌های کاربردی که نحوه استفاده کاربران از وب‌سایت را بیان می‌کند، برای کشف دانش از رفتار پیمایشی کاربران می‌باشد. گام‌های اصلی در کاربرد کاوی وب پیش‌پردازش داده‌ها، کشف الگو و آنالیز الگو می‌باشد. در مرحله پیش‌پردازش، داده‌های جمع‌آوری شده به فرم داده مورد نیاز برای مرحله کشف الگو تبدیل می‌شود. در مرحله کشف الگو از تکنیک‌های مختلف داده‌کاوی مانند خوشه‌بندی، قوانین انجمنی ، تحلیل آماری و الگوهای دنباله‌ای برای کشف الگوهای رفتاری کاربران و بدست آوردن اولویت‌ها و علائق کاربر استفاده می‌شود. در مرحله آنالیز الگو، الگوهای نامربوط از میان الگوهای بدست آمده حذف

[1] Web Mining
[2] Linkage

می‌گردد. الگوهای بدست آمده از فرایند کاربرد کاوی می‌تواند در حیطه‌های مختلف مانند شخصی‌سازی، تجارت الکترونیک، بهبود وب‌سایت و واکشی اولیه و حافظه نهان بکار گرفته شود. در سال‌های اخیر، تحقیقات در حوزه‌ی وب‌کاوی و به خصوص در زمینه کاربردکاوی وب گسترش یافته است. از اواسط ۱۹۹۰ تا به حال هزاران مقاله در زمینه‌ی وب‌کاوی منتشر شده است که حدودا پنجاه درصد از این مقالات درباره‌ی کاربردکاوی وب بوده است.

هنگام مواجهه با مسئله‌ای در کاربردکاوی وب، مشکل اصلی انتخاب مناسب‌ترین تکنیک برای مسئله مورد بحث است. مانند همه زمینه‌های دیگر، برای مسائل کاربردکاوی وب پاسخ یکتا وجود ندارد. در مطالعات صورت گرفته، از میان تکنیک‌های مختلف داده‌کاوی، تمرکز اصلی بر تکنیک خوشه‌بندی می‌باشد. در کاربرد کاوی وب خوشه‌بندی برای گروه‌بندی اشیائی که مشخصات یکسان دارند، مورد استفاده قرار می‌گیرد. خوشه‌بندی در کاربرد کاوی وب به خوشه‌بندی صفحات وب و خوشه‌بندی کاربران وب تقسیم می‌شود. خوشه‌بندی صفحات وب صفحات وب اصول ذاتی صفحات وب را انعکاس می‌دهد، در حالیکه خوشه‌بندی کاربران وب علاقه گردشگری کاربران را نشان می‌دهد. هدف اصلی در این پژوهش مدل کردن رفتار گردشگری کاربران براساس خوشه‌بندی کاربران علاقه‌مند و نمایش فضای وظایف گردشگری در سطح معنایی، از طریق وب‌کاوی و تحلیل معنایی پنهان احتمالاتی می‌باشد.

در خوشه‌بندی کاربران وب، کاربرانی که هنگام گردشگری در وب‌سایت رفتارهای مشابهی انجام می‌دهند در یک گروه قرار می‌گیرند. هدف نهایی خوشه‌بندی کاربران همانند دیگر تکنیک‌های داده‌کاوی، فراهم کردن قابلیت تحلیل خوشه‌ها به منظور استخراج هوش تجاری یا استفاده از آن‌ها برای اعمالی نظیر شخصی‌سازی وب و بهبود وب‌سایت‌ها و واکشی اولیه صفحات و غیره می‌باشد. بسیاری از مطالعات صورت گرفته در زمینه خوشه‌بندی کاربران وب، برای مدل کردن و گروه‌بندی کاربران وب از تکنیک‌های استاندارد خوشه‌بندی استفاده می‌کند. اما الگوهای کاربردی بدست آمده از این تکنیک‌ها، الگوهایی در سطح صفحات وب می‌باشند. این الگوها نه تنها مشخصات ذاتی فعالیت‌های کاربران را بدست نمی‌آورند، بلکه فاکتورهای اساسی و غیرقابل مشاهده‌ای را که منجر به الگوهای گردشگری خاص می‌شود تعیین نمی‌کنند.

بنابراین برای درک بهتر عواملی که منجر به الگوهای گردشگری می‌شود نیاز به توسعه و گسترش تکنیک‌ها می‌باشد، بطوریکه تکنیک‌های جدید بتوانند بصورت خودکار هدف‌های اصلی گردشگری کاربران را مشخص کند و رابطه معنایی میان کاربران و همچنین میان اشیاء وب را تعیین کنند. مدل‌های متغیر پنهان، همانند تحلیل معنایی پنهان احتمالاتی[1] بصورت گسترده‌ای برای کشف رابطه پنهان میان هم‌رخدادی اشیاء بکار می‌رود. تحلیل معنایی پنهان احتمالاتی روش احتمالاتی را برای کشف متغیرهای پنهان ارائه می‌دهد که روشی بسیار انعطاف‌پذیر و دارای پایه آماری بسیار قوی می‌باشد. اساس تکنیک تحلیل معنایی پنهان احتمالاتی مدلی به نام مدل aspect می‌باشد. فرض کنید مجموعه‌ای از عامل‌های پنهان وجود دارد که اساس هم‌رخدادی میان دو مجموعه از اشیاء می‌باشد، تکنیک تحلیل معنایی پنهان احتمالاتی از الگوریتم ماکزیمم انتظار وقوع برای برآورد مقادیر احتمالاتی که رابطه بین عامل‌های پنهان و دو مجموعه از اشیاء اندازه‌گیری می‌کند، استفاده می‌کند. دانش پس‌زمینه در تکنیک تحلیل معنایی پنهان احتمالاتی مفهوم عامل پنهان می‌باشد که فرض می‌کند عامل‌های پنهان قطعیت هم‌رخدادی کاربر و صفحات را کنترل کند. تحلیل معنایی پنهان بطور

[1] Probabilistic Semantic Latent Analysis

موفقیت‌آمیزی در کاربرد کاوی وب معرفی شده است و موفقیت بزرگی را در مطالعات مربوطه مانند مدلسازی کاربر، سیستم‌های پیشنهاددهنده و غیره بدست آورده است.

در این پژوهش، کاربرد کاوی وب براساس تحلیل معنایی پنهان احتمالاتی ارائه می‌شود. در کاربرد کاوی وب داده‌های هم‌رخداد کاربران وب و اشیاء وب می‌باشند. در این پروژه فاکتورهای پنهان که رابطه میان موجودیت‌ها را نشان می‌دهد به عنوان وظیفه[1] شناخته می‌شود. این مطلب برای تایید این حقیقت است که فاکتورهای پنهان اهداف گردشگری کاربران در وب‌سایت‌ها و همچنین تعاملاتشان با اشیاء وب را نشان می‌دهد. با اعمال تکنیک تحلیل معنایی پنهان احتمالاتی، می‌توان بطور موثر مشخصات عامل‌های پنهان را شناسایی کرد، سپس از لحاظ کمی رابطه بین کاربران وب و وظایف همچنین رابطه بین اشیاء وب و کاربران وب را اندازه‌گیری کرد. این روابط بصورت احتمالاتی اندازه‌گیری می‌شود که به نوبه خود به کشف الگوهای کاربردی مختلف با استفاده از استنتاج احتمالاتی منجر می‌شود.

۱-۲ هدف پژوهش

در این پروژه مطالعات صورت گرفته بر کشف الگوهای کاربردی از طریق کاربرد کاوی وب و سپس استفاده از الگوهای کشف شده برای ایجاد گروه‌های کاربران علاقه‌مند و استفاده از گروه‌های بدست‌آمده برای کاربرد واکشی اولیه صفحات می‌باشد. هدف اصلی در این پژوهش در روشی که از ساختار سلسله مراتبی وب‌سایت بجای صفحات وب استفاده می‌شود، بهبود در مدل کاربر و در نتیجه بهبود در گروه‌بندی کاربران وب می‌باشد که از تکنیک تحلیل معنایی پنهان احتمالاتی در کاربرد کاوی وب استفاده می‌کند. همچنین کاربرد گروه‌های ایجاد شده برای واکشی اولیه صفحات ارائه می‌شود.

۱-۳ ساختار پژوهش

ساختار پژوهش به صورت زیر می‌باشد. فصل دو به ارائه‌ی اطلاعات پیش‌زمینه‌ای لازم در مورد کاربرد کاوی وب تخصیص پیدا می‌کند. مرور بر ادبیاتی که در این فصل آمده است برای دادن اطلاعات به خواننده برای درک موضوع است.

فصل دو به جزئیات وب کاوی و بطور اخص، کاربرد کاوی وب می‌پردازد. با استفاده از رویکرد بالا به پایین از وب کاوی شروع کرده و سپس بحث را اختصاصی کرده و به کاربرد کاوی وب پرداخته می‌شود.

فصل سه، به معرفی تکنیک خوشه‌بندی کاربران وب پرداخته می‌شود. این فصل شامل دو بخش است، در بخش اول کارهای انجام در زمینه خوشه‌بندی کاربران وب با استفاده از تکنیک‌های داده‌کاوی بررسی می‌شود. در بخش دوم، کارهای صورت گرفته با استفاده از تکنیک‌های آنالیز معنایی در زمینه خوشه‌بندی کاربران وب معرفی و بررسی می‌شود.

فصل چهارم به ارائه روش پیشنهادی برای خوشه‌بندی کاربران وب بر اساس تحلیل معنایی پنهان احتمالاتی پرداخته می‌شود که فصل اصلی این در پژوهش می‌باشد. بخش بعد جزئیات مربوط به پیاده سازی، بررسی نتایج آزمایشات و مقایسه و تحلیل آنها با سایر روش‌ها مطرح می‌شود.

[1] Task

فصل پنجم به ارزیابی روش پیشنهادی اختصاص دارد. در این فصل نتایج آزمایشاتی که برای ارزیابی سیستم انجام شده است، ارائه خواهد شد و رفتار الگوریتم تفسیر می‌شود. در بخش دوم این فصل نیز الگوریتم پیشنهادی برای وظیفه حقیقی واکشی اولیه صفحات اعمال می‌شود تا درخواست‌های آینده کاربران گروه‌بندی شده بر طبق صفحات عمومی مشاهده شود توسط کاربران، پیش‌بینی شود و نتایج بدست آمده با روش‌های دیگر مقایسه شده می‌شود.

و نهایتا فصل ششم که به جمع‌بندی و خلاصه‌کردن نتایج فصل‌های پیشین اختصاص دارد، و دیدی کلی بر مسئله ایجاد می‌کند.

فصل دوم

استخراج اطلاعات از داده‌های کاربرد وب

۲-۱ مقدمه

در این فصل به تعریف مفاهیم در زمینه وب‌کاوی خواهیم پرداخت، بعد از ارائه‌ی دیدی کلی بر مسئله‌ی وب‌کاوی، به کاربردکاوی‌وب خواهیم پرداخت. روش‌های موجود برای این نوع کاوش در ادبیات این مسئله، در این فصل عنوان شده و کاربردهای آن ذکر می شود.

۲-۲ داده کاوی و استخراج دانش

با گسترش سیستم‌های پایگاه داده‌ای و حجم بالای داده‌های ذخیره شده در این سیستم‌ها، برای اینکه بتوان داده‌های ذخیره شده را پردازش کرد نیاز به ابزارهایی است که بتواند اطلاعات حاصل از این پردازش را در اختیار کاربران قراردهد. با استفاده از پرس و جوهای ساده در SQL و ابزارهای گوناگون گزارش‌گیری معمولی، می‌توان اطلاعاتی را در اختیار کاربران قرارداد تا با استفاده از آن بتوانند به نتیجه‌گیری در مورد داده‌ها و روابط منطقی میان آنها بپردازند. اما وقتی که حجم داده‌ها بالا باشد کاربران هرچند زبردست و مجرب باشند نمی‌توانند الگوهای مفید را در میان حجم انبوه داده‌ها تشخیص دهند و حتی اگر قادر به اینکار هم باشند هزینه عملیات از نظر نیروی انسانی و مادی بسیار بالا است .

علاوه بر این، با وجود سیستم‌های یکپارچه اطلاعاتی، سیستم‌های یکپارچه بانکی و تجارت الکترونیک ، لحظه به لحظه به حجم داده‌ها به پایگاه داده‌های مربوطه اضافه شده و باعث بوجود آمدن انبارهای عظیمی از داده‌ها شده‌است بطوری که ضرورت کشف و استخراج سریع و دقیق دانش از این پایگاه داده‌ها را بیش از پیش نمایان کرده‌است.

داده‌کاوی بعنوان یک راه حل برای این مسائل مطرح می‌باشد در متون علمی تعاریف گوناگونی برای داده کاوی ارائه شده است برخی از این تعاریف عبارتند از:

- داده‌کاوی عبارت است از فرایند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه داده‌های بزرگ و استفاده از آن در تصمیم‌گیری در فعالیت‌های تجاری مهم.
- اصطلاح داده‌کاوی به فرایند نیمه‌خودکار تجزیه و تحلیل پایگاه داده‌های بزرگ به منظور یافتن الگوهای مفید اطلاق می‌شود.
- داده‌کاوی یعنی جستجو در پایگاه داده‌ها برای یافتن الگوهایی میان داده‌ها.
- داده‌کاوی یعنی تجزیه و تحلیل مجموعه داده‌های قابل مشاهده برای یافتن روابط مطمئن بین داده‌ها.

کاوش اطلاعات، حجم عظیمی از داده‌های خام را به فرمی تغییر می‌دهد که انسان بتواند آنها را به راحتی بفهمد و برای تصمیم‌گیری بتواند از این اطلاعات استفاده‌کند. درکل، داده‌کاوی به دو شاخه ی مرتبط مدل کردن و کشف الگو تقسیم می‌شود. مدل کردن خلاصه کردن ساختارهای عظیم داده‌است. مثال‌هایی از چنین مدل کردنی عبارتند از:

- تحلیل جزءبندی
- تحلیل رگرسیون
- تجزیه‌ی سری‌های زمانی

و فرآیند کشف دانش، می‌تواند در قدمهای زیر خلاصه شود[۱]:

- پاک‌سازی داده
- انتخاب و تغییرشکل داده
- کاوش‌داده
- ارزیابی
- ارائه‌ی دانش

دو قدم اول، فاز پیش‌پردازش و دو قدم آخر فاز پس‌پردازش هستند. قدم کاوش‌داده، قسمتی از فرآیند کشف دانش است که از داده‌های خام دانش استخراج می‌کند. در این قدم از استراتژی‌های یادگیری‌ماشین، آمار و ... برای حل مسائل داده‌کاوی استفاده می‌شود. دانش استخراج شده باید دارای صفات دقیق، قابل درک، جالب باشد [۲].

البته درجه‌ی اهمیت هر کدام از این صفات بسته به کاربر و کاربرد دارد. برای مثال اگر قرار باشد کاربر از دانش استخراج شده استفاده کند و تصمیم‌گیری‌های استراتژیک توسط کاربر انجام شود، قابل درک بودن دانش استخراج شده اهمیت بیشتری پیدا می‌کند. و همچنین اگر دانش استخراج شده قابل فهم نباشد، هدف تحقق پیدا نخواهدکرد.

در مسائل داده‌کاوی، هر چه حجم داده‌ها بیشتر می‌شود، میل بیشتری برای کشف الگوهای مخفی در داده‌ها برای به دست آوردن سود تجاری و ... به وجود می‌آید. در قدم اصلی داده‌کاوی ممکن است از چندین الگوریتم داده‌کاوی استفاده شود. کار اصلی الگوریتم داده‌کاوی با توجه به نوع مسئله‌ی کشف دانش تغییر می‌کند.

دو نوع اصلی الگوریتم‌های داده‌کاوی، کلاس‌بندی و خوشه‌بندی می‌باشد. خوشه‌بندی به فرایند تقسیم‌بندی داده به یک یا چند گروه به طوری که فاصله بین خوشه‌ها حداکثر و فاصله درون خوشه‌ها حداقل باشد، اطلاق می‌شود. با توجه تعریف خوشه‌بندی، در این پروژه الگوریتم مورد استفاده و پیشنهادی در دسته‌ی خوشه‌بندی قرار می‌گیرد، چرا که در نهایت کاربران‌وب به تعداد مشخصی گروه تقسیم‌بندی می‌شوند به این ترتیب که کاربرانی که در مواجهه با وب‌سایت رفتار گردشگری مشابهی دارند، در یک گروه قرار می‌گیرند.

وب رسانه‌ای است برای رسیدن به اطلاعات زیاد و متنوعی که در سایت‌های مختلف در شبکه گسترده جهانی ذخیره شده‌است. رشد گسترده وب مسبب رشد ثابت اطلاعات شده‌است که چندین مشکل را به همراه دارد، بدست آوردن اطلاعات مرتبط، استخراج دانش مفید و یادگیری در مورد عملکردهای مشتریان یا کاربران، نمونه‌هایی از این مشکلات

است. بنابراین در داده‌کاوی حوزه‌ای به نام وب‌کاوی بوجود آمده است که سعی بر حل این مشکلات دارد. در بخش بعدی در مورد وب‌کاوی و انواع آن بحث می‌شود.

۲-۳ وب‌کاوی[1]

وب، اکنون بزرگترین انبار داده‌ی بشر است. ایجاد وب برمی‌گردد به سال ۱۹۹۰ که یک کارمند CERN اولین برنامه برای حرکت روی اتصالات دو طرفه بین مجموعه‌های از اسناد را با رابط‌کاربرگرافیکی[2] نوشت. نام این نرم‌افزار را وب‌گسترده‌ی جهانی[3] گذاشتند. اگرچه عمل هدایت بین اسناد به این طریق، جدید نبود ولی اضافه شدن رابط کاربر کار جدیدی بود. این رابط کاربر جدید با ابداع دیگری درCERN به هم آمیخت. این ابداع HTTP[4] بود. در سال 1993حجم ترافیک HTTP ده برابر شد [۳] و بعد از آن وب گسترده‌ی جهانی از اسم یک نرم‌افزار تبدیل شد به یک وب واقعی که با ارتباطات بین اسناد ابرمتن به وجود آمده‌است.

وب‌کاوی استفاده از تکنیک‌های داده کاوی برای کشف و استخراج خودکار اطلاعات از مستندات و سرویس‌های وب می‌باشد. در واقع وب‌کاوی به تمام فرایند کشف اطلاعات گفته می‌شود و نه صرفا نرم‌افزارهای کاربردی که از ابزارهای استاندارد داده‌کاوی استفاده می‌کنند. برخی از صاحب‌نظران در این حوزه وظایف وب‌کاوی را به چهار وظیفه زیر تقسیم می کنند که در ادامه آورده می‌شود:

- پیدا کردن منبع : یعنی فرایند بازیابی داده‌ها از اسناد وب مورد نظر.
- انتخاب و پیش‌پردازش اطلاعات: در این مرحله به صورت خودکار اطلاعات خاصی از اسناد بازیابی شده، انتخاب و پیش پردازش می شوند.
- تعمیم [۴] **:** در این مرحله به صورت خودکار الگوهای عام در یک یا چندین سایت وب کشف می‌شود. برای این عمل تکنیک‌های مختلف داده‌کاوی، یادگیری ماشین و متدهای مبتنی بر وب استفاده می‌شود.
- تحلیل**:** در این مرحله الگوهای به دست آمده در مرحله قبل اعتبار سنجی [۵] و تفسیر می‌شوند.

از طرفی عواملی که بر درک و ارزیابی کاربر از صفحات وب در هنگام فرایند داده‌کاوی تاثیر می‌گذارد عبارتند از:

الف) محتوای صفحه وب
ب) طراحی صفحه وب
ج) طراحی کل سایت که شامل ساختار آن می باشد.

عامل اول در ارتباط با داده‌هایی است که توسط یک سایت ارائه می‌شود. عامل دوم در ارتباط با روشی است که محتویات سایت برای کاربران در دسترس و قابل درک می‌شود. توجه شود که میان طراحی یک صفحه‌ی مستقل و طراحی کل یک سایت فرق وجود دارد، زیرا یک سایت شبکه‌ای است از صفحات مرتبط است و تا وقتی که طراحی یک سایت ساده و قابل درک نباشد، کاربران تمایل به استفاده از وب سایت ندارند.

---
[1] Web mining
[2] GUI
[3] Word Wide Web
[4] HyperText Transfer Protocol

۲-۴ چالش‌های وب‌کاوی

وب‌کاوی با چالش‌ها و محدودیت‌های متنوعی روبه‌رو است. از یک دیدگاه می‌توان این محدودیت‌ها را به دو گروه تکنیکی و غیرتکنیکی تقسیم کرد. از محدودیت‌های غیرتکنیکی می‌توان به عدم پشتیبانی مدیریت، کافی نبودن بودجه و عدم وجود منابع مورد نیاز مانند نیروی انسانی متخصص اشاره‌کرد. اما مشکلات تکنیکی بسیار است که به برخی از آنها در این جا اشاره می‌شود:

۱. **داده‌های ناصحیح و نادقیق:** برای آن که فرآیند وب‌کاوی با موفقیت انجام شود، لازم است داده‌های جمع آوری شده صحیح و در قالب مناسب باشند. اما معمولا مشکلات زیادی در این زمینه وجود دارد، چرا که ممکن است داده‌ها دقیق نباشند و از طرفی داده‌ها می‌توانند ناکامل بوده و برخی مقادیر موجود نباشد. همچنین تخمین میزان اطمینان درباره صحت و دقت داده‌ها به سادگی امکان‌پذیر نیست.
۲. **عدم وجود ابزارها:** محدودیت دیگر وب‌کاوی، عدم وجود ابزارهای مناسب و کامل برای آن می‌باشد. متخصصان باید تصمیم بگیرند آیا برای یک کاربرد وب‌کاوی، ابزار خاص آن را توسعه دهند و یا از ابزارهای موجود استفاده‌کنند.
۳. **ابزارهای سفارشی:** ابزارهای موجود تنها یکی از انواع وب‌کاوی مانند طبقه‌بندی یا خوشه‌بندی را پشتیبانی می‌کنند. اما بهتر آن است که یک ابزار قادر به انجام چندین تکنیک وب‌کاوی باشد تا کاربران بتوانند با توجه به نیازمندی‌های خود از تکنیک مناسب استفاده کنند.

البته در حال حاضر تحقیقات بسیاری در زمینه وب‌کاوی در حال انجام است که هدف آن‌ها حل این مشکلات می‌باشد.

۲-۵ اجزای اصلی وب‌کاوی

همان طور که گفته شد وب منبع عظیمی از داده است. این داده یا محتویات وب است که از طریق میلیون‌ها صفحه‌ی وب به دست می‌آید و یا اطلاعات اتصالات بین صفحات وب و یا داده‌های دسترسی روزانه کاربران به صفحات وب که در سرورهای وب در فایل‌های متنی به نام ثبت[1] ذخیره می‌شوند. براساس داده‌های مورد استفاده در کاوش وب، سه شاخه‌ی مجزا در وب‌کاوی به وجود آمده‌است. این سه شاخه، عبارتند از:

۱- محتواکاوی وب[2]
۲- ساختارکاوی وب[3]
۳- کاربردکاوی وب[4]

[1] Log
[2] Web content mining
[3] Wes structure mining
[4] Web usage mining

۲-۵-۱ محتواکاوی وب

استفاده از تکنیک‌های داده‌کاوی به منظور استخراج اطلاعات مفید از محتوای صفحات وب، محتواکاوی وب نامیده می‌شود. این اطلاعات عبارتند از متن ، تصویر ، ویدئو و ... همچنین استخراج منابع وب ، طبقه‌بندی مستندات ، دسته‌بندی و استخراج اطلاعات از صفحات وب بخشی از وظایف محتواکاوی‌وب است. در اینترنت به دلیل عدم وجود ساختاری برای مستندات، عملیات مکانیزه کردن استخراج، سازماندهی و مدیریت اطلاعات وب مشکل است. کاربردهای معمول این نوع کاوش وب، سازماندهی وب بر اساس محتوا و درجه بندی[1] صفحات وب بر اساس محتوا است.

محتواکاوی وب، از جهاتی شبیه به داده کاوی و متن‌کاوی است. به داده‌کاوی شبیه است زیرا تکنیک‌های داده کاوی زیادی در مسئله‌ی محتواکاوی وب استفاده می‌شوند و به متن‌کاوی شبیه است زیرا اکثر محتویات وب، متن هستند. ولی از جهاتی هم با داده‌کاوی متفاوت است زیرا داده‌های وب اکثراً نیمه ساختاریافته هستند در حالیکه داده کاوی با داده‌های ساختاریافته سروکار دارد. با متن‌کاوی نیز از آن جهت متفاوت است که متن موجود در وب نیمه ساخت‌یافته است ولی متن‌کاوی با داده‌های متنی بدون ساختار سروکار دارد.

۲-۵-۲ ساختارکاوی وب

ساختارکاوی وب ساختار فرامتن را پیدا می‌کند . در واقع ساختارکاوی وب به کشف مدل پس زمینه حاکم بر ساختار فراپیوندهای وب می‌پردازد و هدف آن ایجاد اطلاعاتی همچون تشابه یا ارتباط بین سایت‌های مختلف وب است. ساختارکاوی وب درگیر تحلیل لینک‌های داخلی و خارجی صفحات وب است و برای رتبه‌بندی موتور جستجو به کار می‌رود. این بخش سعی در کشف ساختارهای لینک‌های پشت صفحات وب می‌کند که اطلاعات بیشتری از مستندات مفید را بیان می‌کنند. همچنین ساختارکاوی وب، عمل کشف اعتبار سایتها، برای موضوعات و بازنگری سایت‌هایی است که به چندین شخص حقیقی یا حقوقی وابسته هستند. لینک‌های درون یک صفحه، عمومی بودن یک صفحه را مشخص می‌کند در حالی که لینک‌های بیرونی یک صفحه توانگری صفحه را نشان می‌دهد و یا یک صفحه‌ای که بصورت دائم تکرار می‌شود می‌تواند یک صفحه مهم باشد. بنابرین می‌توان از میان ساختارهای لینک، وب را کاوش کرد . Page Rank، CLEVER مثالهایی برای ساختارکاوی وب است.

۲-۵-۳ کاربردکاوی وب

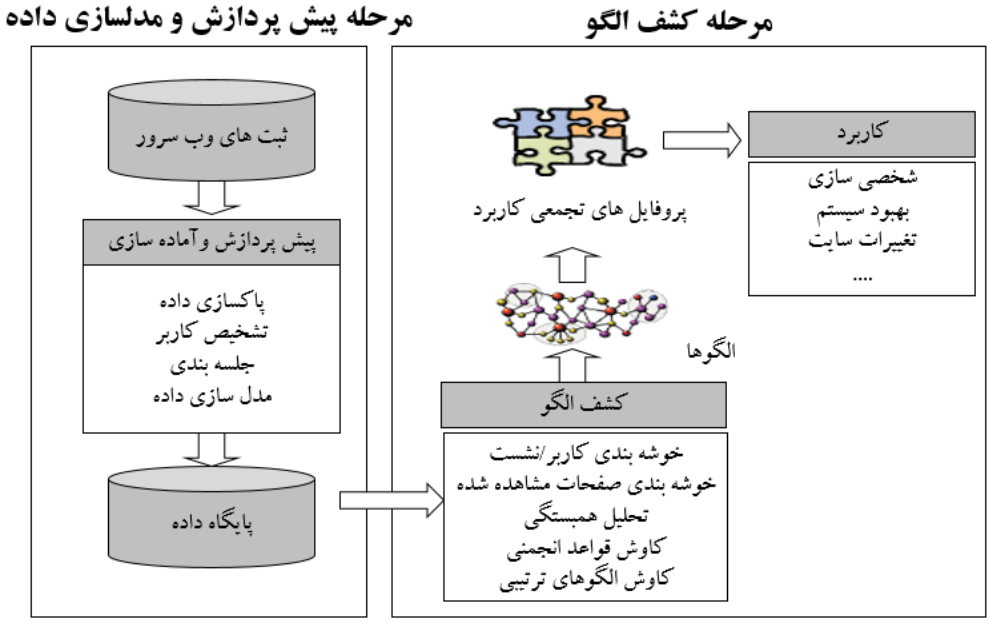
کاربردکاوی وب بخشی از وب‌کاوی است که استخراج مکانیزه الگوی دسترسی کاربر از یک یا چند سرویس‌دهنده وب را بر عهده دارد. کاربردکاوی وب سعی می‌کند از تعاملات کاربر با وب اطلاعاتی کسب کند و از آنها بصورت سابقه‌ای در مراجعات بعدی کاربر سود ببرد. داده‌های مورد استفاده وب شامل داده‌های ثبت‌شده ، User Session ، User Profile ، Browser Log ، Proxy Server Log ، Web Server usage Log ، User Query ، Cookies ، User Transaction ، Bookmark Log و .... می‌شوند. سازمان‌ها اغلب در عملیات روزانه‌شان حجم زیادی از اطلاعات را جمع‌آوری و تولید می‌کنند. بیشتر این اطلاعات معمولا بطور خودکار توسط سرویس‌دهنده‌های وب تولید و توسعه Server Access Log ، جمع‌آوری می‌شوند. سایر اطلاعات کاربر شامل Refer Log ها است که شامل اطلاعاتی درباره صفحات استنادی برای منشأ صفحه و ثبت‌نام کاربر با ابزارهای کلی جمع‌آوری داده مانند CGI است . تحلیل Server

[1] Ranking

Access Log و اطلاعات ثبت‌شـده کاربران می توانند اطلاعات ارزشمندی در مورد اینکه چگونه بایستی یک وب سایت را در یک سازمان ایجاد کرد فراهم کنند.

از آنجایی که موضـوع اصـلی پروژه کاربردکاوی‌وب می‌باشـد، در ادامه این بخش به ارائه‌ی مسـائل مرتبط با این نوع کاوش وب پرداخته می‌شـود. فرآیند کلی کاربردکاوی وب شـامل سه مرحله است: آماده سازی و مدلسازی داده، کشف الگو از داده های کاربرد وب و اسـتفاده از الگوهای کشـف شده برای کاربردهایی مانند شخصی‌سازی وب، واکشی اولیه صـفحات،تجارت الکترونیک. در مرحله آماده سـازی داده، پیش‌پردازش‌های لازم بر روی ثبت‌های جمع‌آوری‌شده، بیان می‌شـود. ثبت‌های وب برای مقاصـد مختلف حذف می‌شـوند، مثلا برای مرتب کردن داده‌های غیرلازم (مثلا دسـترسی از طریق web spider)، برای تعیین نشست‌های کاربران[1] (به وسیله‌ی کوکی[2]ها)، برای ذخیره‌ی داده‌ها در یک پایگاه‌داده‌ی رابطه‌ای و در نهایت ثبت های خام وب به فرمتی تبدیل می‌شـود که می‌تواند در داده‌کاوی مورد اسـتفاده قرار گیرد. در مرحله کشـف الگو وظیفه های مختلف داده‌کاوی مانند خوشـه بندی، کاوش قواعد انجمنی و کشـف الگوهای ترتیبی را می‌توان بر روی داده‌ی پیش پردازش شـده اجرا کرد. نتایج فاز کاوش به پروفایل های تجمعی کاربرد تبدیل می‌شوند که پروفایل‌های بدسـت آمده در این مرحله برای کاربردهایی مانند واکشـی اولیه صـفحات، شـخصـی‌سـازی وب ، تغییردر وب‌سایت و تجارت الکترونیک .

در شـکل ۲-۱ یک نمای جامع از فرآیند کاربردکاوی وب ارائه داده می‌شـود. در ادامه‌ی این بخش از این چارچوب نشـان داده شـده از شـکل ۲-۱ به عنوان راهنما اسـتفاده خواهد شـد. و در ادامه بحث مفصلی از فعالیت های کاربردکاوی وب لازم برای این فرآیند شـامل پیش پردازش، تکنیک های متداول کشـف الگو که بر روی این داده های کاربرد بکار می‌روند ارائه خواهد شد.



شکل ۲-۱ فرایند کاربردکاوی وب [۸]

[1] User sessions
[2] Cookie

۲-۶ آماده‌سازی و مدل‌سازی داده

یک مرحله‌ی مهم در هر عملیات داده‌کاوی ایجاد مجموعه داده‌ی مناسبی است که بر روی آن الگوریتم های داده‌کاوی اجرا می‌شوند. این فرایند می‌تواند شامل پیش‌پردازش داده‌های اولیه، تبدیل داده‌ها به فرمتی مناسب به عنوان ورودی عملیات داده‌کاوی، باشد. به مجموعه‌ی این عملیات آماده‌سازی داده گفته می‌شود.

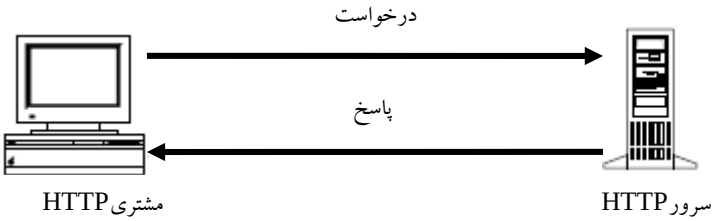
فرایند آماده‌سازی داده اغلب پرهزینه‌ترین مرحله از نظر زمان و محاسبات در فرآیند کشف دانش محسوب می‌شود [۶] و در مورد کاربردکاوی وب نیز این موضوع صادق است. مرحله‌ی آماده‌سازی داده در وب‌کاوی کاربرد وب اغلب نیازمند استفاده از الگوریتم‌های خاص و مکاشفه‌ای است که در حوزه‌های دیگر داده‌کاوی در این مرحله بکار برده نمی‌شوند. این مرحله در استخراج موفقیت‌آمیز الگوهای مفید از داده‌ها بسیار حیاتی است. در ادامه موضوعات و مفاهیم مرتبط با آماده‌سازی و مدل سازی داده‌ها در کاربردکاوی وب توضیح داده می‌شود.

۲-۶-۱ منابع داده

منابع عمده‌ی داده در کاربردکاوی وب فایل‌های ثبت وب‌سرور می‌باشد که شامل ثبت‌های دسترسی وب‌سرور است.

داده های کاربرد

پیش ازآنکه به داده‌های کاربرد و منابع آن پرداخته شود ابتدا تعاریفی از متاداده‌هایی که توسط وب‌سرورها تولید و استفاده می‌شوند ارائه می‌گردد [۷] . شکل۲-۲ یک تراکنش HTTP[1] را بین یک مشتری HTTP و یک سرور HTTP نشان می‌دهد. برای سادگی فرض کنید که مشتری HTTP یک مشتری وب است و یک سرور HTTP نیز یک وب‌سرور می‌باشد. یک مشتری وب که برای کاربران انسانی طراحی شده‌است یک مرورگر وب نامیده می‌شود مانند Mozilla Firefox، Netscape Navigator و Microsoft Internet Explorer . مثال‌هایی از وب‌سرور عبارتند از IBM HTTP Server، Apache HTTP Server و Microsoft Internet Information Server (IIS).



شکل ۲-۲ تراکنش HTTP

در یک تراکنش HTTP داده‌های کاربرد[2] با متاداده‌های زیر تعریف می‌شوند:

- آدرس IP ماشین مشتری
- شناسه کاربر
- زمانی که سرور پردازش درخواست را انجام می‌دهد.

[1] HyperText Transfer Protocol
[2] Usage data