روشی جهت بهینه‌سازی نتایج موتورهای جستجوگر با بهره‌مندی از الگوریتم‌های خوشه‌بندی و بردار پشتیبان

محمداحمدی
دانشجوی کارشناسی ارشد- کامپیوتر نرم‌افزار- موسسه آموزش عالی غیرانتفاعی-غیردولتی مقدس اردبیلی

عباس میرزایی
گروه مهندسی کامپیوتر- واحد اردبیل- دانشگاه آزاد اسلامی- اردبیل- ایران

چکیده

امروزه دانش و اطلاعات وب در حال افزایش است، همچنین تعداد وب‌سایت‌ها نیز روز به روز بیشتر می‌شود؛ در نتیجه نیاز است شرایطی فراهم گردد تا در بحث‌هایی مثل تجارت الکترونیک، موتورهای جستجو، تبلیغات هدفمند و غیره بر اساس رویکرد کاربران مبتدی و حرفه‌ای از بین صفحات مختلف پیشنهاد مناسبی به آنها داده شود. اما این‌که بتوان نیازها و علایق کاربران مختلف را پیش‌بینی نموده و صفحات مورد علاقه آنان را پیشنهاد کرد با پیچیدگی‌هایی همراه است و باید بتوان با روش‌هایی بهینه و با استفاده از الگوهای حرکتی کاربران که در سرورهای وب ثبت گردیده است این کار را انجام داد. روش‌های مختلفی برای پیش‌بینی صفحات وب ارائه شده است. در این تحقیق روشی جهت بهبود عملکرد موتورهای جستجو با استفاده از خوشه‌بندی و بردار پشتیبان ارائه شده است که از چند گام مختلف اعم از پیش‌پردازش داده، ایجاد پروفایل ثبت کاربر، استخراج میزان علاقه‌مندی کاربر، خوشه-بندی و بردار پشتیبان تشکیل شده است. ارائه روش پیشنهادی علاوه بر کاهش میزان حجم اطلاعات، پیچیدگی سیستم را کم نموده است، همچنین در مرحله پیش‌بینی، با ایجاد تغییراتی در مدل بردار پشتیبان میزان دقت سیستم نیز نسبت به مدل پایه افزایش یافته است. مدل بردار پشتیبان و روش خوشه‌بندی استفاده شده در این سیستم به صورت بهینه عمل کرده تا نتایج قابل قبولی از سیستم استخراج گردد. معیارهای ارزیابی در این تحقیق دقت، فراخوانی و پوشش می‌باشد. میزان دقت کلی سیستم برابر با ۸۳/۵ شده است کهنسبت به کارهای مشابه ۱/۸ بهبود حاصل شده است. نتایج ارزیابی این سیستم به کمک ابزار متلب با دیگر وش‌های مشابه بیانگر عملکرد بهتر و بهینه است.

واژگان کلیدی: وب‌کاوی، خوشه‌بندی، الگوهای رفتاری کاربران، بردار پشتیبان

مقدمه

موتورهای جستجو امروز همچون قلب تپندهٔ اینترنت عمل می‌کنند و عملاً با افزایش سرسام‌آور تعداد صفحات موجود در وِب، تنها راه کاوش و یافتنِ اطلاعات در این شبکه، استفاده از موتورهای جستجو می‌باشد. امروزه کمپانی‌هایی که موتورهای جستجوی خود را عرضه کرده‌اند رقابتی سنگین و فشرده را در عرصهٔ عمومی آغاز کرده‌اند که نمونهٔ آن رقابت شرکت‌های گوگل و یاهو می‌باشد که بتازگی مایکروسافت نیز موجودیت خود را در این عرصه شروع کرده است و قصد سرمایه‌گذاری وسیع در این حوزه را دارد. شرکت گوگل با الگوریتم انقلابی خود که البته پیشینه‌اش در کارهای موتورهای قبلی به چشم می‌خورد در حال حاضر محبوب‌ترین موتور جستجو می‌باشد و سهام آن ارزشی افسانه‌ای یافته است و این موفقیت مدیون نگرش جدید مؤسسین این موتور به ماهیت مسئله می‌باشد. موتورهای سنتی با الگوریتم‌هایی که مبتنی بر شمارش کلمات کلیدی است پیش می‌روند و این نوع نگرش به تنهایی جواب‌گویِ نیازهای ذهنی مراجعه‌کنندگان به موتورهای جستجو نیست. در ارائهٔ پاسخ‌ها الگوریتم‌های رتبه‌بندی[1] مطابق خواست کاربر عمل نمی‌کنند و نمونه‌های کم و غیر تجاری‌ای از موتورهای جستجو وجود دارد که به مسئلهٔ شخصی‌سازی[2] جستجو اهمیت داده‌اند. لذا با توجه به اهمیت مسئله، ضرورت مطرح گشتن روش‌هایی نوین در این حوزه آشکارتر می‌گردد. رشد رو به گسترش اینترنت و به تبع آن، توسعه کسب‌وکارهای الکترونیک در جهان باعث شده تا وبسایت‌ها از اهمیت بالایی برخوردار شده، نقش غیرقابل انکاری در برقراری ارتباط الکترونیکی بین سازمان‌ها و مؤسسات با مشتریانشان پیدا کنند؛ چرا که در چنین محیط الکترونیکی‌ای، وب‌سایت، پل ارتباطی بین سازمان‌ها و مشتریانشان بوده، مشتریان از طریق صفحات مختلف وب‌سایت به خدمات سازمان دسترسی پیدا می‌کنند. در سال‌های ابتدای توسعه اینترنت و کسب‌وکارهای الکترونیک، سازمان‌ها و شرکت‌ها درصدد توسعه هرچه بیشتر وب‌سایت، افزایش صفحات وب و جذب هرچه بیشتر بازدیدکنندگان به سایت خود بودند، لکن در سال‌های اخیر، بهبود کیفیت وب‌سایت، اولویت بالاتری پیدا کرده است. در توجیه این مسأله به طور ساده می‌توان گفت که افزایش صفحات وب‌سایت یک سازمان، یک کلاف بزرگ و پیچ‌درپیچ می‌سازد که حرکت در آن و یافتن خدمات یا محصولات مورد نظر را کاری وقت‌گیر، خسته‌کننده و حتی ناموفق می‌کند. این مسأله باعث می‌شود تا سازمان نتواند تمام محصولات و خدماتش را به نمایش بگذارد و در نتیجه احتمالاً مشتری خود را از دست بدهد (Jaiganesh,et al, 2020).

از این‌رو موتورهای جست‌وجو در عصر حاضر نقش کلیدی برای دسترسی کاربران به اطلاعات موجود در فضای وب ایفا می‌کنند. مهم‌ترین وظیفه این سرویس‌ها یافتن بهترین نتایج در زمانی مناسب برای حجم بالایی از درخواست‌های همزمان است. یکی از مهم‌ترین چالش‌های موتورهای جست‌وجو کاستن از زمان پاسخگویی به پرس‌وجوهای دریافتی است که این مسئله مستقیما بر میزان رضایتمندی کاربران و درآمد موتور جستجو تاثیر گذار است. بنابراین استفاده از تکنیک‌هایی که بتواند با توجه به نیازهای کاربران داده‌های متناسب و مطلوبی ارائه دهد یک برتری و مزیت محسوب می‌شود لذا یکی از روش‌های پرکاربرد در زمینه افزایش رضایت‌مندی کاربران و بهبود نتایج مطلوب موتورهای جست‌وجو استفاده از وب‌کاوی و آنالیز رفتار کاربران گذشته است.

بیان مساله

در عصر نوین با انبوه درخواست‌ها و نیازهایی که کاربران به کمک موتورهای جست‌وجو درصدد مرتفع کردنشان هستند استفاده از تکنیک‌های سریع و کارا یک چالش و ضرورت سرویس‌دهندگان است. بجهت رفع مشکل رتبه‌بندی در جواب‌های یک پرس‌وجویِ ارائه شده به یک موتور جستجو می‌توان با استفاده از اطلاعات بدست آمده از کاربر به یکی از دو طریقِ استخراج از زمینهٔ جستجو یا استفاده از پروفایلِ کاربر، جواب‌های اولیه را دوباره بردار‌بندی کرد و آنگاه نتایج این بردار‌بندی را با استفاده از یک الگوریتمِ خوشه‌بندی، دسته‌بندی کرد و آنگاه با اعمال الگوریتم‌های مارکوف به یکه زیرمجموعهٔ کوچک از جواب‌ها رسید که بهبود قابل توجهی نسبت به جواب‌هایی که موتورهای جستجویِ متداول به کاربران می‌دهند را در بر دارد. رتبه‌بندی جدیدی از نتایج پرس‌وجو بدست آورد که در عمل بسیار به خواست کاربر نزدیکتر باشند و شرایطی جدید را در تنوع مشخصات ارائه شده نیز برآورده کند موتورهای جستجویِ کنونی در عمل با

[1] Ranking
[2] personalization

دادنِ نتایجی با شمار زیاد کاربر را سردرگم می‌کنند و هدف این الگوریتم و شمایِ کلی آن است که به یک ساختارِ جدید در ارائهٔ نتایج به کاربر دست یابد و او را از مواجه شدن با نتایجی با شمار بیش از اندازه‌ای که بتواند آن‌ها را بسنجد باز دارد(Jaiganesh,et al, 2020). تعداد نتایجی که یک کاربر تخصصیِ گوگل پس از تحویلِ پرس‌وجویش می‌آزماید بندرت از تعداد انگشتانِ دست فراتر می‌رود. و همین مسئله ضرورت موتورهائی فراگیر و با این قابلیت جدید را بیش از پیش گوشزد می‌کند. پروفایل کاربر با روشی جدید که مبتنی بر یادگیری از رویِ تاریخچهٔ جستجوهای کاربر است ارائه می‌گردد و این زمانی است که این پروفایل به روشی ساده‌تر مانند انتخاب مستقیم خود کاربر از مجموعه‌ای محدود از پروفایل‌ها، ممکن نباشد. نتایج اولیه از موتورِ جستجو با رتبه‌بندی بهبود نیافته نیز گرفته شده‌اند و اکنون باید ابتدا با استفاده از بردارِ پروفایل، بردار بندی بر رویِ نتایج موجود اولیه اعمال شود. روشی که در بردار‌بندی بکار می‌رود باید بتواند در اعمال علایق کاربر به نتایج، کارآ باشد و نتیجهٔ کلی الگوریتم را توجیه کند. روش خوشه‌بندی نیز باید بتواند در توجیه خود بنابر معیارهای کیفیت خوشه‌بندی به مقدار قابل قبولی دست یابد و سرانجام اینکه رتبه‌بندی ثانویه باید بهبود قابل توجه‌ی نسبت به رتبه‌بندی اولیه از خود نشان دهد که این بهبود باید در راستای علایق کاربر و شخصی‌سازیِ نتایج جستجو باشد و در ضمن کاربر را از افتادن در دامِ جواب‌هایی با رتبه‌بندیِ بالا اما یکدست و شبیه به هم برهاند و یک تنوع و کیفیت را در دادن زیرمجموعهٔ کوچکی از جواب‌هایِ اولیه رعایت کند (Robertson, et al, 2019)

هر موتور جستجوئی ناگزیر از آن است که قضاوت و رتبه‌بندیِ نهایی خود را نه بر اساس چند کلمهٔ محدود یا براساس ارجاعات صفحات به همدیگر (روش مورد استفاده در گوگل)، که بر اساس نیازهای کاربری که به اطلاعات نیاز دارد، انجام دهد و قطعاً به پروفایلی از کاربر نیاز دارد. با توجه به پیشرفت‌های روزافزون در فناوری دیسک‌های سخت قاعدتاً می‌توان برای هر کاربر یک یا چند پروفایل بسته نوع کاربر ذخیره کرد و آنگاه در هر بار ورود کاربر به محیط جست‌وجو بر اساس پروفایلش که پویا بروز می‌شود جست‌وجو را انجام داد. امکان بروز رسانی پویای کاربر بنظر ناگزیر می‌رسد تا بتوان همواره علایق کاربر را ردگیری کرد. به نظر می‌رسد که استفاده از تکنیک‌های وب-کاوی در این فرایند ناگزیر باشد و به جهت دستیابی به سرعت و دقتی قابل قبول این تکنیک‌ها می‌توانند کمک شایانی در رتبه‌بندی جواب‌های ارائه شده به کاربر داشته باشند. در این میان الگوریتم‌های خوشه‌بندی، بردار پشتیبان و رتبه‌بندی صفحات با کاربرد چند گانه می‌توانند نقش اصلی را ایفا کنند. در روش ارائه شده در این تحقیق از مدل بردار پشتیبان استفاده شده است و برای کاهش پیچیدگی از پروفایل کاربران و ایجاد انجمن‌ها استفاده شده است. در حقیقت پروفایل کاربران شامل اطلاعات مربوط به زمان و توالی بازدید صفحات کاربران و میزان علاقه‌مندی کاربران در یک نشست می‌باشد. برای افزایش دقت سیستم مورد ارائه همچنین از الگوریتم خوشه‌بندی k-means به صورت تلفیقی با مدل بردار پشتیبان استفاده شده است.

اهمیت موضوع

با توجه به افزایش حجم وب، پیمایش و جستجوی آن از اهمیت بالایی برخوردار است. در پیمایش این حجم وسیع از صفحات بهتر آن است، صفحاتی ابتدا پیمایش شوند که مرتبط با موضوع موردنظر می‌باشند."پیمایش هوشمند" روشی است که برای پیمایش صفحات مرتبط با یک موضوع به کار می‌رود. در این روش سعی بر آن است که در هنگام پیمایش، صفحات هاب[1] (صفحاتی است که به صفحات مهمی اشاره می‌کنند) خوب تشخیص داده شوند تا از آن‌ها به عنوان منبعی برای رسیدن به اعتبار صفحه یا اتوریته[2] (صفحاتی هستند که محتوای مهمی دارند) استفاده شود (Okabe and Seiji, 2017). وب طی یک فرآیند آشفته و غیر متمرکز در حال رشد است و این روند منجر به تولید حجم وسیعی از مستندات متصل به یکدیگر گشته است که از هیچ گونه سازمانده‌ی منطقی برخوردار نیستند. در واقع وب به مجموعه بزرگی از داده‌های ساخت یافته و نیمه ساخت یافته تبدیل شده است که کاربران آن از همپوشانی داده‌ها رنج می-برند. بنابراین تحلیل رفتارهای کاوشی کاربران وب و بررسی واقعی علایق کاربران اهمیت خاصی پیدا کرده است.

[1] Hub
[2] Page Authority

رشد چشم‌گیر اطلاعات در وب جهان گستر به طور ناگهانی منجر به اضافه بار اطلاعاتی شده است و از این رو پیدا کردن یک بخش خاص از اطلاعات دشوار و وقت‌گیر است. سیستم‌های شخصی‌سازی در سال‌های اخیر به منظور مقابله با این مشکل و با هدف ارائه یک تجربه شخصی به کاربران بر اساس اولویت‌ها، و نیازهای فردی آنها ایجاد شده‌اند. با رشد نمایی صفحات وب و پدیده گردآوری نمودن اطلاعات ضرورت بهره‌گیری از یک موتور جستجوی قدرتمند که به کاربر بتواند پیشنهادهای مناسب را ارائه بدهد، ملاحظه می‌شود. در این راستا کارهای متعددی انجام پذیرفته است. برای این منظور بعضا از الگوریتم‌های فراابتکاری و برخی دیگر از مدل‌های تصادفی بهره برده‌اند. بنابراین ارائه روش‌های بهینه‌سازی جهت بهبود نتایج موتورهای جستجوگر یک الزام در این حوزه محسوب می‌شود.

ادبیات و پیشینه

کاظمی در سال ۱۳۹۳، برای افزایش دقت در حذف کردن حالات و داشتن دقت بهتر در پیش بینی مدل، روشی را برای حذف بر اساس خطا ارایه می‌کند که براساس این روش برای حذف یک حالت با مرتبه بالاتر به تخمین خطای حالت‌های مرتبه پایین تر با مجموعه‌ای از جلسه‌هایی که توسط مرتبه بالاتر هم پوشش داده می‌شود، می‌پردازد. در واقع در این روش برای هر حالت به تعداد حالات مرتبه بالاتر آن خطا محاسبه می‌شود.

مقیمی و همکارانش در سال ۱۳۹۳، ابتدا فایل رخداد سرویس‌دهنده وب را به کمک روش‌های پیش پردازش به فایلی حاوی نشست‌های کاربر تبدیل می‌کند و خروجی مرحله اول را که با استفاده از مدل خوشه بندی کامیانه بهینه شده، خوشه‌بندی می‌کند، سپس این مجموعه را توسط مدل مارکوف همه مراتب براساس دو روش متفاوت مورد بررسی قرار داده و مرتبه‌ای که بالاترین صحت کار را می‌دهد بعنوان برچسب برای آن خوشه انتخاب می‌کند. آخرین مرحله هم ورود فرد جدید به سایت و پیش‌بینی بهترین گزینه برای وی می‌باشد.

Chintan و همکارانش در سال ۲۰۱۹، مروری بر آنچه در داده‌کاوی وب وجود دارد صورت دادند که شامل فرآیند وب کاوی کاربردی[1] و بررسی عمیق کارهایی که تا کنون در روش پیش پردازش داده‌ها برای وب کاوی کاربردی انجام شده است. هنگامی که کاربران با وب سایت ارتباط برقرار می‌کنند آثاری با فرمت‌های مختلف در مکان‌های مختلف تولید و برجای می‌گذارند. این آثار در راه مناسب ثبت و ضبط می‌شوند، سیاهه‌ها ممکن است از ناخالصی‌ها و اختلالات جمع‌آوری می‌شوند، از اینرو تکنیک‌های مختلف داده‌کاوی نمی‌تواند به طور مستقیم بر آنها اعمال شود. بنابراین نیازمندی‌ها، مراحل و روش‌های پیش پردازش داده‌ها شرح داده شده است. تکنیک‌های مختلف الگوی کشف که می‌تواند در پیش‌پردازش سیاهه‌ها جمع‌آوری شود در مرحله قبل به کار برده شود و به استخراج دانش از آن منجر گردد، شرح داده می‌شود.

Priyanka و همکارانش در سال ۲۰۲۰، فرایند استخراج از فایل‌های سرور وب به منظور استخراج الگوهای استفاده شده در پیش‌بینی لینک‌های وب با کمک مدل مارکوف ارائه شده است. نتیجه این روش پیش‌بینی صفحه وب محبوب و یا جهت‌دهی رفتار کاربر می‌باشد. روش پیشنهادی جهت‌دهی رفتار کاربر تشخیص جامعه می‌شود بر اساس اندازه‌گیری شباهت دو به دو آن‌ها همراه با مدل مارکوف با مفهوم الگوریتم آپریووری که برای لینک وب استفاده می‌شود.

Meera و همکارش در سال ۲۰۱۵، یک چارچوب دو لایه‌ای را ارائه کرده‌اند. این چارچوب دو لایه ای از دو لایه به صورت آفلاین و برخط تشکیل شده است. بدین صورت که لایه‌ی اول مربوط به مرحله‌ی آموزش می‌باشد و از اطلاعات کاربران در فاز آفلاین بهره گرفته می‌شود و لایه‌ی دوم مربوط به فاز برخط می‌باشد که عملیات پیش‌بینی در این فاز به نمایش گذاشته می‌شود. در این مقاله از مدل‌های مارکوف و مارکوف مخفی به منظور مدل‌های پیش‌بینی کننده استفاده شده است نتایج هر کدام از مدل‌ها با استفاده از تئوری ترکیب شواهد دمپستر- شافر ترکیب شده و صفحاتی که احتمال بازدید آن‌ها بیشتر است نمایش داده می‌شود. این چارچوب قابل تغییر می‌باشد و از مدل‌های دیگر همچون خوشه‌بندی، قوانین انجمنی و غیره می‌توان در آن استفاده کرد و به آن افزود. خروجی مبتنی بر تلفیق مارکوف و مارکوف مخفی نشان می‌دهد که سیستم از دقت بالاتری نسبت به تک تک هر کدام از مدل‌ها دارا می‌باشد.

[1] Web Usage Mining (WUM)

Madhuri و همکارانش در سال ۲۰۱۹، روشی جهت رشد نامحدود خدمات وب و سیستم‌های اطلاعاتی مبتنی بر وب، حجم داده کلیک-استریم[1] و داده کاربر جمع‌آوری شده توسط سازمان‌های مبتنی بر وب در فعالیت‌های روزانه خود به حجم بالایی رسیده است. تحلیل چنین حجم بالایی از داده می‌تواند به ارزیابی اثربخشی کمپین‌های تبلیغاتی، بهینه‌سازی عملکرد کاربردهای مبتنی بروب کمک کند و محتوای سفارش‌گیری را برای ویزیتورها فراهم کند. در کار قبلی، ما روشی پیشنهاد کردیم، تحلیل الگوی رابطه‌ای گری با استفاده از زنجیره‌های مارکوف، که شامل کشف الگوهای معنی‌دار و ارتباطات از یک مجموعه بزرگ داده است، که اغلب در لاگ‌های دسترسی به سرور برنامه کاربردی و وب، لاگ‌های پروکسی[2] و غیره ذخیره شده‌اند. در اینجا ما یک رویکرد جدید برای تحلیل رفتار گشت‌وگذار کار با استفاده از GRPA[3] با زنجیره‌های مارکوف با طول متغیر پیشنهاد می‌کنیم. یکVLMC[4]یک گسترش مدل است که به تاریخچه طول متغیر اجازه ضبط شدن می‌دهدGPRA با زنجیره‌های مارکوف با طول متغیر بطور موثر و اثربخشی نشاندهنده اطلاعات متوالی در داده استفاده از وب است و می‌تواند بسط داده شود تا باعث ادغام با یک مدل پیش‌بینی رفتار گشت‌وگذار کاربر وب برای کاربردهای استخراج استفاده از وب بهتر شود. تعداد زیادی از صفحات وب روی بسیاری از وبسایت‌ها باعث مشکلات گشت‌وگذار می‌شوند. زنجیره‌های مارکوف اخیرا برای مدلسازی رفتار حرکتی کاربر روی وب جهان‌گستر استفاده می‌شوند.

Zhu و همکارانش در سال ۲۰۱۸، روشی را برای ایجاد یک مدل مارکوف از وبسایت‌های مبتنی بر رفتار ویزیتوری گذشته پیشنهاد کرده-اند که از مدل مارکوف برای ایجاد پیش‌بینی‌های لینک استفاده شده که به کاربران جدید برای گشت‌وگذار در وبسایت‌ها کمک می‌کنند. یک الگوریتم برای فشرده‌سازی ماتریس احتمال تراکنش استفاده می‌شود تا صفحات وب با رفتارهای تراکنشی مشابه را کلاستربندی کند و ماتریس تراکنش را به یک اندازه بهینه برای محاسبه احتمال اثربخشی در پیش‌بینی لینک فشرده کند. یک روش فوروارد مسیر حداکثری برای اصلاحات بیشتر درمورد اثربخشی پیش‌بینی لینک استفاده می‌شود. پیش‌بینی لینک در یک سیستم برخط به نام جستجوگر مرور برخط اجرا می‌شود تا به حرکت کاربر در وب‌سایت انطباقی کمک کند.

Montgomery و همکارانش در سال ۲۰۲۰، نشان داده‌اند که چطور اطلاعات مسیر می‌تواند با استفاده از مدل احتمالی چندگانه مرور وب گروه‌بندی و تحلیل شود. در روش ارائه شده با استفاده از داده‌ها از فروشگاه برخط اصلی ارزیابی شده است. نتایج نشان می‌دهند که مولفه حافظه این مدل در پیشگویی صحیح یک مسیر ضروری است. در مقایسه، مدل‌های ماکوف مرتبه اول و احتمالی چندگانه سنتی مسیرها را به طرز ضعیفی پیش‌بینی می‌کنند. این نتایج پیشنهاد می‌کنند که مسیرها ممکن است اهداف یک کاربر را منعکس کنند، که می‌تواند در پیش‌بینی حرکات آینده در یک وب‌سایت مفید باشد. یک کاربرد احتمالی مدل تحقیق پیش‌بینی تبدیل خرید است. ما متوجه می‌شویم که تنها بعد از شش مشاهده می‌توان با دقت بیشاز 40%خریداران را پیش‌بینی کرد که بسیار بهتر از نرخ پیش‌بینی تبدیل خرید 7%بنچ مارک انجام شده بدون اطلاعات مسیر است. این روش می‌تواند برای شخصی‌سازی طراحی وب و پیشنهاد محصول بر اساس مسیر کاربر استفاده شود.

Kalbhor و همکارانش در سال ۲۰۱۹، تلاش کرده‌اند تا با استفاده از اطلاعات در خصوص صفحات از قبل بازدید شده کاربر و مرور سوابق آن، مجموعه بعدی صفحات وبی که ممکن است مشاهده نماید را پیش‌بینی نموده و مشکل طبقه‌بندی را برطرف نمایند. این برای پیش‌بینی رفتار کاربر بسیار مفید است در حالی که او از اینترنت به دلایل بسیاری مانند افزایش سرعت مرور و یا به حداقل رساندن بار سرور و غیره استفاده می‌نماید. در این مرجع بر اساس مدل پنهان مارکوف فازی پیش‌بینی روند بعدی کاربر پیشنهاد شده است که از نسخه k امین مرتبه مارکوف استفاده می‌نماید. این مرجع همچنین گزارشی از مقایسه روش‌های مختلف پیش‌بینی درخواست آینده با روش مناسب خود ارائه می‌نماید.

Yang و همکاران در سال ۲۰۱۸، به بررسی الگوریتم‌های فیلترینگ مشارکتی به کار گرفته شده در برنامه‌های کاربردی اینترنتی تلفن همراه پرداخته شده است. در ابتدا یک چارچوب برای سیستم پیشنهاد دهنده بر اساس داده‌های کاربران مختلف از جمله رتبه‌بندی کاربر

[1] Click Stream
[2] Proxy log
[3] Grey Relational Pattern Analysis
[4] Variable Length Markov Chains

و رفتارهای کاربران پیشنهاد می‌شود. سپس ویژگی‌های کلیدی این دو نوع داده‌ها بحث شده است. علاوه بر این، چندین نوع الگوریتم به عنوان روش‌های مبتنی بر حافظه و روش‌های مبتنی بر مدل طبقه‌بندی و مقایسه شده است و در پایان دو مطالعه موردی در مجموعه داده‌های مووی لینز بر اساس رتبه‌بندی کاربر بر اساس رفتار کاربر برای تعیین اعتبار این چارچوب ارائه شده است.

Katarya و همکارشدر سال ۲۰۱۷، یک سیستم توصیه‌گر ارائه داده‌اند که علاقه کاربران را نسبت به مجموعه‌ای از آیتم‌ها (کتاب، فیلم، موزیک، نرم‌افزار) جمع‌آوری می‌کند. این اطلاعات به صورت صریح (معمولا به صورت جمع‌آوری امتیازان) و یا ضمنی (به وسیله رصد فعالیت‌های کاربران از قبیل آهنگ‌های گوش داده، نرم‌افزارهای دانلود شده، صفحات وب بازدید شده و غیره) به دست می‌آیند. این سیستم‌های توصیه‌گر تلاش می‌کند که به وسیله پیشنهاد اطلاعاتی که مورد علاقه کاربر باشد بر مشکل سرریز اطلاعات غلبه کند. با افزایش محبوبیت شبکه‌های اجتماعی، اطلاعات شبکه اجتماعی رشد سریعی داشته است و باعث شده است که کاربران با سر ریز اطلاعات مواجه می‌شوند.

اهداف و فرضیه‌های پژوهش

از جمله اهداف مهمی که در این تحقیق می‌توان به آن‌ها اشاره کرد عبارتنداز:

هدف اصلی:

- بهینه‌سازی نتایج موتورهای جستجوگر با بهره‌مندی از الگوریتم‌های خوشه بندی و بردار پشتیبان

اهداف فرعی :

۱) افزایش دقت موتورهای جستجو با بهره‌گیری از الگوی رفتاری کاربران
۲) بهره‌گیری از بردار پشتیبان در داخل خوشه‌ها به منظور افزایش دقت سیستم پیشنهادی
۳) ایجاد پایگاه داده آموزشی قوی از روی الگوی رفتاری کاربران جهت تخمین رفتارهای آتی کاربران
۴) ارائه یک موتور جستجو بر وب براساس الگوهای رفتاری ایجاد شده از کاربران و الگوریتم بردار پشتیبان

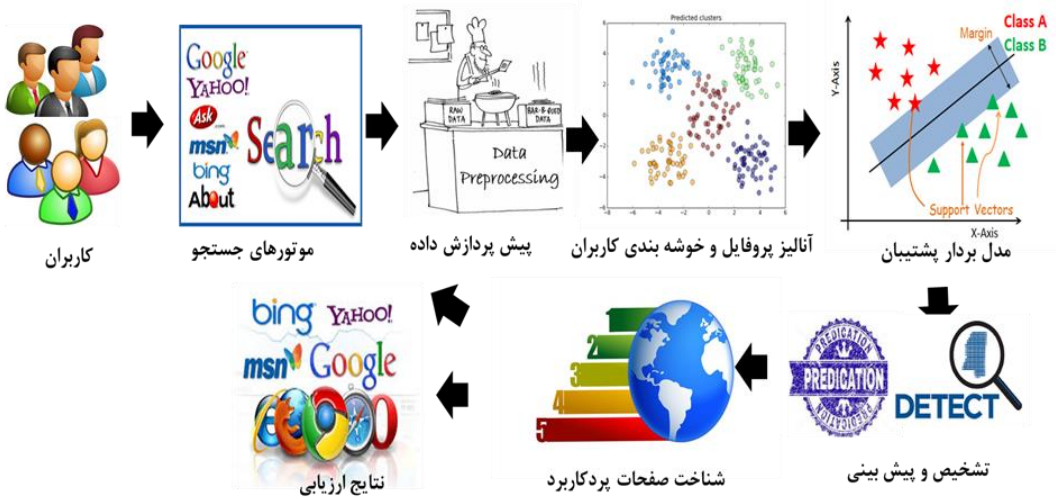
فرضیه‌های تحقیق عبارتند از:

۱) می‌توان با بهره‌گیری از بردار پشتیبان و خوشه‌بندی اقدام به ارایه توصیه مناسب در موتورهای جست‌وجو نمود.
۲) می‌توان با بهره‌گیری از فایل‌های ثبت کاربران امکان ایجاد یک پروفایل برای آنها براساس الگوی رفتاری آنها ایجاد نمود.
۳) بهره‌گیری از بردار پشتیبان منجر به افزایش دقت و کاهش زمان پاسخ سیستم پیشنهادی می‌گردد.

روش تحقیق

روش پیشنهادی بر اساس وب‌کاوی مبتنی بر کاربرد ارائه شده است که می‌تواند بدون نیاز به دخالت کاربران و به صورت کاملا صریح، نیازها و علاقه آنان را تشخیص داده و پیشنهادهای مناسب را به آن‌ها ارائه نماید. تشخیص الگوهای رفتاری کاربر و استخراج اطلاعات با ارزش از بین این میزان داده گسترده شامل چندین فاز می‌باشد که عبارتنداز: پالایش داده و پیش پردازش، تبدیل داده، استفاده از تکنیک‌های داده‌کاوی و تفسیر نتایج. در این تحقیق برروی ایجاد پروفایل کاربران تمرکز شده است. در ادامه بر روی یادگیری بردار پشتیبان تمرکز شده است و برای مدل کردن پیش‌بینی رفتار پیمایشی کاربر بر روی وب مناسب عمل می‌نمایند. به صورت کلی ورودی برای این مسائل دنباله‌ای از صفحات وب است که به وسیله کاربر ملاقات شده و نتیجه ساختن داده آموزشی است که بتواند محتمل‌ترین صفحه‌ای که کاربر به عنوان صفحه بعدی خواهد دید را پیش‌بینی کند. در این تحقیق از بردار پشتیبان استفاده شده و برای کاهش پیچیدگی از پروفایل کاربران استفاده شده است. در حقیقت پروفایل کاربران شامل اطلاعات مربوط به زمان و توالی بازدید صفحات کاربران و میزان علاقه‌مندی کاربران در یک نشست می‌باشد. برای افزایش دقت سیستم مورد ارائه همچنین از الگوریتم خوشه-بندی به صورت تلفیقی با SVM استفاده شده است.

در این روش، تعاملات کاربران وب‌سایت مورد نظر که شامل پروفایل کاربر، صفحات وب و اطلاعات ثبت شده مربوط به استفاده کاربران می‌باشد، مورد پردازش قرار می‌گیرند. کاربران با جست‌وجوهای مختلف از طریق موتورهای جستجو رفتارشان ثبت و ذخیره می‌شود. این رفتار و تعاملات در قالب پروفایل کاربری که برگرفته از شاخص‌های مهمی مانند: رفتار، زمان، علائق و غیره است ذخیره می‌شود. این پروفایل‌ها باید مورد آنالیز و بررسی قرار بگیرند تا به کمک تکنیک‌های مختلف صفحات مورد بازدید و نیاز کاربران شناخته شوند. در ادامه ابتدا معماری روش پیشنهادی مطابق با شکل (۱) تشریح شده و سپس گام‌های تعریف شده برای رسیدن به هدف تحقیق با جزئیات لازم تشریح شده است همچنین به چگونگی ارزیابی و شبیه‌سازی اشاره شده است.



شکل ۱. بلوک دیاگرام روش پیشنهادی

پیش پردازش داده

در این مرحله، ساختار وب‌سایت و محتوای آن مورد پردازش قرار گرفته و یک مدل از آن تولید می‌گردد. یک وب سایت، شامل تعداد زیادی صفحات وب است که در آن، صفحات وب از طریق پیوندها به هم مرتبط‌اند. ابتدا نیاز است که داده را از فایل‌های ثبت وقایع متعدد یکسان‌سازی کنیم؛ تا کاربران یکسان را در فایل‌های ثبت وقایع متفاوت تعیین شوند. این فایل‌ها می‌توانند به فرمت‌های معمولی[1](آدرس IP کاربر، زمان و تاریخ دستیابی،روش درخواست (Get, Post,...)، آدرس صفحه دستیابی شده، پروتکل (HTTP HTPP1.1)، کد خروجی و تعداد بایت‌های جا به جا شده) یا توسعه یافته[2] ذخیره شوند. در فرمت توسعه یافته فیلد ارجاع دهنده[3] و فیلدهای عامل کاربر[4] وجود دارند. اولین گام در پیش‌پردازش داده حذف محتوای نامرتبط است. برای مثال فایل‌های تصویری مانند gif jpg, با چک کردن رشته‌ی رکوردها در فایل ثبت وقایع حذف می‌شوند. اگر این رکوردها توسط جست‌وگر وب ایجاد شود با مقایسه با robot.txt حذف می‌شود.

ایجاد پروفایل برای کاربران و ترسیم ماتریس کاربر –صفحه

مرحله‌ی تشخیص کاربران پس از پاک‌سازی داده‌ها آغاز می‌شود. تشخیص کاربران یک مرحله‌ی بسار پیچیده است. از آنجا که ممکن است فایل ثبت وقایع کاربران توسط یک سرور وب تنها ضبط شود و یا ممکن است توسط پروکسی در ترکیب با سرورهای وب پیچیده ضبط شود. در این گام، آدرس یا DNS آدرس ارجاع دهنده، آدرس الکترونیک و عامل کاربرد، زمینه‌های مرتبطی در فایل ثبت وقایع

[1] CLF (common log file)
[2] ELF(extend log file)
[3] Referrer url
[4] User agent

هستند. آدرس DNS که همان IP است در CLF[1] قرار دارد. آدرس ارجاع دهنده[2]، کاربر تایید شده[3]، عامل کاربردی[4] در فایل ELF[5] قرار دارند. در طول تعاملات کاربران با صفحات وب‌سایت، تمامی رفتارهای کاربران که شامل مرور صفحات وب‌سایت می‌باشد در یک فایل متنی ذخیره می‌گردند. به منظور رسیدن به الگوهای استفاده از وب سایت، باید فایل‌های ثبت رفتارهای کاربران مورد پردازش قرار گیرد. برای شناسایی کاربران و تشخیص الگوهای استفاده از وب مربوط به آن کاربر با توجه به ساختار فایل‌های ثبت رخداد سمت سرور، راه-کاری ارائه شده که کاربران مختلف را از هم متمایز می‌نماید همچنین الگوهای استفاده از صفحات وب را تشخیص می‌دهد. جهت انجام این کار، در این بخش از مشخصه شناسه نشست استفاده می‌شود اما مشکلی که در این مورد می‌تواند ایجاد شود این است که یک کاربر باید به یک مرورگر و محدوده زمانی خاصی محدود شود و همچنین نمی‌تواند استفاده‌های مکرر یک کاربر را تشخیص دهد یا به عبارت دیگر، درخواست‌های مشاهده صفحات مربوط به یک کاربر، محدود می‌شود به درخواست‌های که در طول مدت یک نشست ارائه شده است. روشی که در این تحقیق در نظر گرفته‌ایم استفاده از آدرس IP به همراه شناسه Agent که دسترسی به هر دو خیلی راحت‌تر بوده و در فایل‌های ثبت رخداد سمت سرور نیز ثبت می‌شوند و مشکلی که در این روش وجود دارد این است که نمی‌تواند منحصر بودن شناسه ترکیبی IP و Agent را تضمین نماید. برای حل این مشکل نیز فرض می‌کنیم که از هر کامپیوتر تنها یک کاربر با یک IP معتبر می‌تواند استفاده نماید. هدف از این مرحله تعیین راه کاری است که بتوان صفحات مشاهده شده توسط کاربران را نشان داد. جهت انجام این کار مجموعه هایی بصورت زیر تعریف می شود:

الف) مجموعه آدرس صفحات وب سایت $P = \{P_1, P_2, \dots, P_n\}$

ب) مجموعه تراکنش های کاربر $T = \{t_1, t_2, \dots, t_m\}$

در اینجا هر تراکنش شامل یک زیر مجموعه از مجموعه P (مجموعه آدرس صفحات) می‌باشد. مطابق با مدل ارائه شده در روش پیشنهادی خوشه‌بندی صفحات وب بر اساس وقوع مشترک در بین نشست‌های کاربران محاسبه می‌شود. برای M نشست وN صفحه یک نشست S_i توسط iامین سطر ماتریس صفحه-نشست می‌تواند نشان داده شود. هر ورودی X_{ik} وزن صفحه P_k در نشست را نشان می-دهد که مقدار ۰ نشان دهنده این است که صفحه P_k در نشست حاضر نیست. بنابراین، در ماتریس صفحه-نشست X هر سطر نشان دهنده یک نشست بر حسب صفحات درخواست شده در آن نشست است.

خوشه بندی بر اساس الگوریتم K-means

داده‌های مهمی در پردازش استفاده از وب وجود دارد که همه آنها ممکن است در دسترس نباشند، که به برخی از آنها مانند فایل‌های ثبت رخداد سمت سرور، محتویات سایت، داده‌های مرتبط با کاربر و غیره می‌توان اشاره کرد. اما مرحله‌ی تشخیص کاربران پس از پاک-سازی داده‌ها آغاز می‌شود. تشخیص کاربران یک مرحله‌ی بسار پیچیده است. از آنجا که ممکن است فایل ثبت وقایع کاربران توسط یک سرور وب تنها ضبط شود و یا ممکن است توسط پروکسی در ترکیب با سرورهای وب پیچیده ضبط شود. در این گام به کمک آنالیز پروفایل کاربران با توجه به میزان علاقه‌مندی در هر نشست، صفحات محبوب انتخاب و در نهایت بر اساس علاقه موجود به صفحات توسط کاربران خوشه‌بندی انجام می‌گردد (Liu, et al, 2017).

برای استخراج میزان علاقه‌مندی کاربران به صفحات در یک نشست از ماتریس میانگین هارمونی به کمک دو معیار فرکانس بازدید صفحه و مدت زمان حضور کاربر در صفحه استفاده می‌شود. رابطه (۱) این مهم را نشان می‌دهد.

$$Interest\ (Page) = \frac{2 * frequency\ (p) * Duration(p)}{freqency(p) * Duration(p)} = z \qquad (1)$$

[1] Common log file
[2] Referrer URL
[3] Authuser
[4] User agent
[5] extend log file

در حقیقت a بیانگرتعداد مراجعات کاربر به یک صفحه وب در یک نشست و b مدت زمان سپری شده بر روی یک صفحه است. این داده‌ها از طریق پروفایل کاربران و بر اساس رفتار کاربران ثبت و دخیره می‌شود. خروجی ماتریس میانگین هارمونیک بصورت ماتریسی از U×P است که سطرهای ماتریس، کاربرها و ستون‌های آن صفحات بازدید شده توسط کاربران را نشان می‌دهد.

$$A_{U\times P} = \begin{array}{c} \\ U1 \\ U2 \\ \vdots \\ Un \end{array} \begin{bmatrix} P1 & P2 & \dots & Pn \\ Z_{11} & Z_{12} & \dots & Z_{1n} \\ Z_{21} & Z_{22} & \dots & Z_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ Z_{n1} & Z_{n2} & \dots & Z_{nn} \end{bmatrix}_{4\times 4}$$

با توجه به تشریح انجام شده برای محاسبه فرکانس صفحه از رابطه (۲) و مدت زمان حضور کاربر در صفحه از رابطه (۳) استفاده می‌شود.

$$a = Frequency\ (u,p) = \frac{number\ of\ visits\ (page)}{\sum_{p\in visited}^{1} p^{(number\ of\ visits(p))}} \quad (۲)$$

$$b = Duration\ (u,p) = \frac{total\frac{duration\ (p)}{size(p)}}{max_{p\in visited(p)}(total\ \frac{duration\ (p)}{size\ (p)}} \quad (۳)$$

در رابطه (۳) Size بیانگر اندازه صفحه بر حسب بایت است. مطابق با روابط تعریف شده هر سطر از ماتریس را می‌توان به صورت برداری بر اساس میزان علاقه هر کاربر به صفحه به صورت $U_۱(a,b,c,\dots)$ و $U_۲(a',b',c',\dots)$ نشان داد.

تشخیص با الگوریتم یادگیری بردار ماشین

از این الگوریتم برای تشخیص صفحات پرکابرد و پیشنهاد به کاربر استفاده می‌شود در حقیقت به کمک شاخص‌های آموزشی این وضعیت آنالیز و بررسی می‌شود و در صورت قابلیت توانایی انجام دادن توسط منبع با اتکا بر حدا آستانه بالا و بهبود جستجوی موتور جستجو انجاممی‌شود. ماشین بردار پشتیبان، الگوریتم طبقه‌بندی[1] بوده و به عنوان یکی از بهترین تکنیک‌های دسته‌بندی و پیش‌بینی و تشخیص outlier شناخته می‌شود و برخلاف الگوریتم‌های خوشه‌بندی در دسته یادگیری با نظارت محسوب می‌شود و دو فاز آموزش و تست دارد. در واقع SVM برای اینکه داده‌های غیر خطی را از هم تفکیک کند باید از کرنل‌های مختلف استفاده کند. برای این کار دیگر در فضای دو بعدی کار نمی‌کند بلکه داده‌ها به فضایی با ابعاد بیشتر نگاشت داده می‌شوند تا بتوان آن‌ها را در این فضای جدید بصورت خطی تفکیک نمود. در حقیقت به کمک این الگوریتم که در دو فاز یادگیری و تشخیص استفاده می‌شود داده‌های آموزشی از پیش به دو کلاس پربازدید و کم بازدید برچسب شده و تقسیم می‌گردند، این دیتاست آموزشی با افزایش آزمایشات تکمیل‌تر از قبل می‌شود و نتایج بدست آمده پشیمانی تخصیص را کاهش می‌دهد.

یافته‌ها

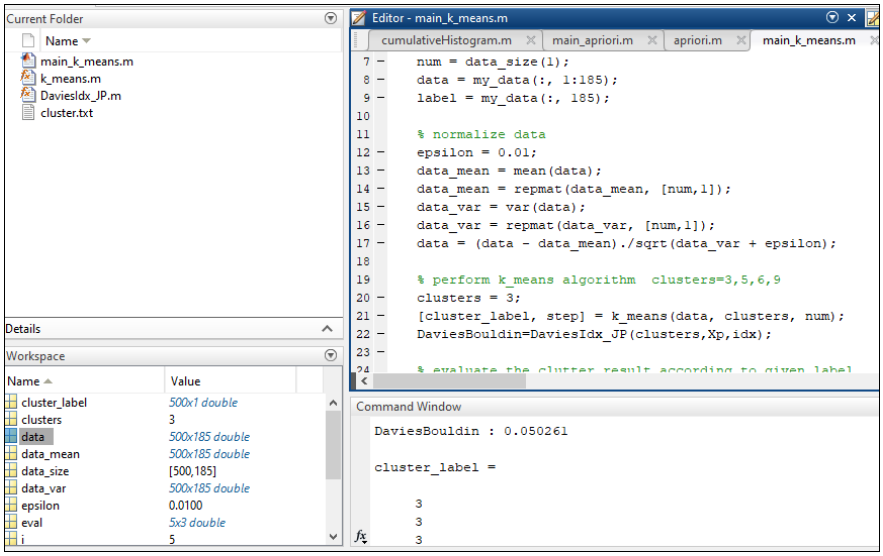
در این تحقیق برای مقایسه و ارزیابی سیستم پیشنهاد دهنده با سایر سیستم‌ها از دیتاست ناسا استفاده شد. این دیتاست در https://www.kaggle.com/datasets/souhagaa/nasa-access-log-dataset دردسترس قرار دارد. ویژگی‌های زیادی در هر رکورد از این دیتاست وجود دارد که شامل آدرس آی پی یا نام دی‌ان‌اسی درخواست‌دهنده تاریخ و ساعت درخواست صفحه آدرس و نام صفحه نوع ارسال صفحه مثلا گت یا پست و غیره بایت ارسالی و دیگر دارد. در اینجا با توجه به حجم زیاد دیتاست نمونه آزمایش شده به تعداد ۵۰۰ رکورد مورد انتخای قرار گرفته است. رکوردهای انتخاب شده از نشست‌های مختلف ثبت شده در زمان‌های مختلف به صورت تصادفی انتخاب شده است. شکل (۲) نمونه‌ای از این مجموعه داده‌ها را نشان می‌دهد.

[1] classifier

```
128.158.41.74 - - [06/Jul/1995:09:37:57 -0400] "GET /icons/menu.xbm HTTP/1.0" 200 527
128.158.41.74 - - [06/Jul/1995:09:37:57 -0400] "GET /icons/image.xbm HTTP/1.0" 200 509
192.189.46.188 - - [06/Jul/1995:09:37:57 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 200 5866
192.189.46.188 - - [06/Jul/1995:09:37:57 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
192.189.46.188 - - [06/Jul/1995:09:37:57 -0400] "GET /images/MOSAIC-logosmall.gif HTTP/1.0" 200 363
192.189.46.188 - - [06/Jul/1995:09:37:58 -0400] "GET /images/USA-logosmall.gif HTTP/1.0" 200 234
herman.and.nl - - [06/Jul/1995:09:37:58 -0400] "GET /pub/winvn/readme.txt HTTP/1.0" 404 -
199.10.135.154 - - [06/Jul/1995:09:37:59 -0400] "GET /icons/menu.xbm HTTP/1.0" 200 527
192.189.46.188 - - [06/Jul/1995:09:37:59 -0400] "GET /images/WORLD-logosmall.gif HTTP/1.0" 200 669
aem003.aeerl.epa.gov - - [06/Jul/1995:09:38:00 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
host62.ascend.interop.eunet.de - - [06/Jul/1995:09:38:00 -0400] "GET /history/history.html HTTP/1.0" 200 1602
www-d2.proxy.aol.com - - [06/Jul/1995:09:38:02 -0400] "GET /shuttle/missions/missions.html HTTP/1.0" 200 8677
199.10.135.154 - - [06/Jul/1995:09:38:04 -0400] "GET /icons/sound.xbm HTTP/1.0" 200 530
jcurro.tiac.net - - [06/Jul/1995:09:38:05 -0400] "GET /shuttle/countdown/liftoff.html HTTP/1.0" 200 4524
163.205.3.60 - - [06/Jul/1995:09:38:06 -0400] "GET /ksc.html HTTP/1.0" 200 7067
163.206.137.21 - - [06/Jul/1995:09:38:07 -0400] "HEAD /shuttle/missions/sts-71/images/KSC-95EC-0874.jpg HTTP/1.0" 200 0
lunatix.bangor.ac.uk - - [06/Jul/1995:09:38:07 -0400] "GET /shuttle/missions/sts-71/images/KSC-95EC-0544.jpg HTTP/1.0" 200 70128
134.217.85.114 - - [06/Jul/1995:09:38:08 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
dd13-031.compuserve.com - - [06/Jul/1995:09:38:09 -0400] "GET /shuttle/missions/sts-71/images/images.html HTTP/1.0" 200 7634
163.205.3.60 - - [06/Jul/1995:09:38:10 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 200 5866
herman.and.nl - - [06/Jul/1995:09:38:10 -0400] "GET /pub/winvn/release.txt HTTP/1.0" 404 -
163.206.137.21 - - [06/Jul/1995:09:38:11 -0400] "GET /shuttle/missions/sts-71/images/KSC-95EC-0875.jpg HTTP/1.0" 200 41160
192.189.46.188 - - [06/Jul/1995:09:38:11 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
aem003.aeerl.epa.gov - - [06/Jul/1995:09:38:12 -0400] "GET /shuttle/resources/orbiters/atlantis.html HTTP/1.0" 200 7025
128.158.41.74 - - [06/Jul/1995:09:38:12 -0400] "GET /history/apollo/apollo-13/images/70HC467.GIF HTTP/1.0" 200 159813
192.189.46.188 - - [06/Jul/1995:09:38:12 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310
aem003.aeerl.epa.gov - - [06/Jul/1995:09:38:13 -0400] "GET /shuttle/resources/orbiters/atlantis-logo.gif HTTP/1.0" 200 4179
163.205.3.60 - - [06/Jul/1995:09:38:13 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
nb1-10.larc.nasa.gov - - [06/Jul/1995:09:38:14 -0400] "GET /shuttle/technology/sts-newsref/stsref-toc.html HTTP/1.0" 200 81920
163.205.3.60 - - [06/Jul/1995:09:38:14 -0400] "GET /images/MOSAIC-logosmall.gif HTTP/1.0" 200 363
163.205.3.60 - - [06/Jul/1995:09:38:15 -0400] "GET /images/USA-logosmall.gif HTTP/1.0" 200 234
163.205.3.60 - - [06/Jul/1995:09:38:15 -0400] "GET /images/WORLD-logosmall.gif HTTP/1.0" 200 669
128.159.124.51 - - [06/Jul/1995:09:38:16 -0400] "GET /ksc.html HTTP/1.0" 200 7067
aem003.aeerl.epa.gov - - [06/Jul/1995:09:38:16 -0400] "GET /shuttle/resources/orbiters/orbiters-logo.gif HTTP/1.0" 200 1932
```

شکل ۲. بخشی از مجموعه داده ناسا مورد مطالعه

با توجه به روش پیشنهادی ارائه شده پس از پیش‌پردازش داده، ایجاد پروفایل ثبت کاربران و استخراج علاقه‌مندی به صفحات در مرحله بعد خوشه‌بندی کاربران جهت کاهش پیچیدگی فضای مساله انجام می‌گردد. قابل ذکر است که مقادیری که صفر نشان داده شده‌اند بیانگر این موضوع است که هیچ نشستی در این صفحه توسط کاربر انجام نشده است. در ادامه با توجه به ۵۰۰ رکورد مورد بررسی از نشست‌های انجام شده ۱۸۵ صفحه مورد آنالیز و بررسی قرار گرفته شده است. بعبارتی کاربران در نشست‌های مختلف میزان علاقه‌مندی متفاوتی به صفحات مورد بازدید داشته‌اند که در اینجا نشان داده شده است. تعداد خوشه‌ها با بررسی چندین وضعیت و انتخاب k بهینه انجام شده است و این امکان به کمک شاخص دیوس بولدین محقق شده است. تعداد خوشه‌های مورد بررسی ۳، ۵، ۷ و ۹ خوشه می‌باشد که از بین سناریوهای مختلف مقدار k=9 انتخاب شده است و این مهم با توجه به کم بودن مقدار شاخص بولدین در مقایسه با سناریوهای دیگر بوده است. در شکل (۳) بخشی از شبیه‌سازی انجام شده برای خوشه‌بندی توسط مدل مورد توسعه در ابزار متلب نشان داده شده است.



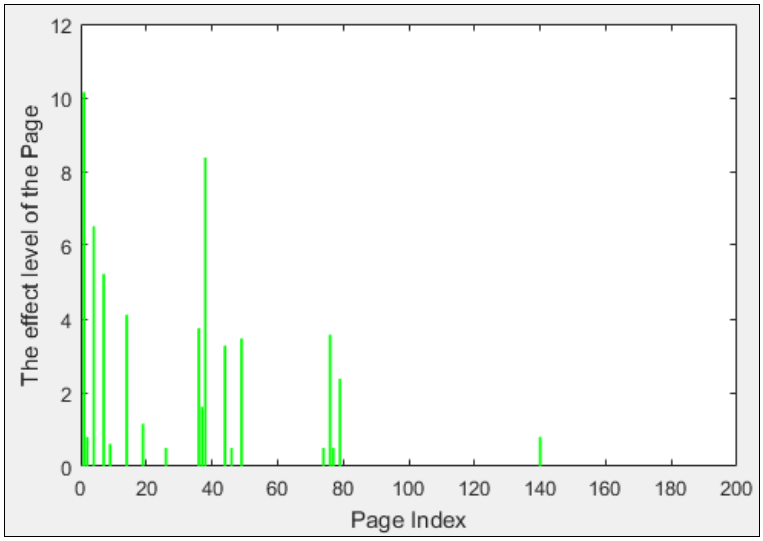
شکل ۳. بخشی از شبیه‌سازی انجام شده برای خوشه‌بندی نشست‌های انجام شده توسط کاربران در ابزار متلب

با توجه به روش پیشنهادی ارائه شده پس از پیش‌پردازش داده، ایجاد پروفایل ثبت کاربران و استخراج علاقه‌مندی به صفحات در مرحله بعد خوشه‌بندی کاربران جهت کاهش پیچیدگی فضای مساله انجام شده است.

جدول ۱. خروجی حاصل از پیش‌بینی کاربران جدید از تعلق به خوشه های تعریف شده

خوشه مورد انتظار	پیش بینی شده		خوشه مورد انتظار	پیش بینی شده
8	'8'		6	'6'
1	'1'		6	'6'
6	'6'		1	'1'
8	'8'		2	'۱'
4	'4'		7	'7'
2	'2'		7	'7'
1	'1'		1	'1'

شکل (۴) میزان علاقه‌مندی کاربران به صفحات در دیتاست مورد مطالعه را نشان می‌دهد. در ادامه هر خوشه پیشنهاد شده با توجه به میزان علاقه‌مندی صفحات موجود در آن خوشه به ترتیب به کاربران پیشنهاد داده می‌شود. در آزمایش انجام شده از بین ۹ خوشه موجود ۶ خوشه انتخاب شده‌اند در حقیقت صفحات موجود در این خوشه‌ها به ترتیب میزان علاقه‌مندی به کاربران مورد تست واقع شده پیشنهاد می‌شود. با توجه به اهمیت صفحات (میزان علاقه‌مندی) در هر خوشه امکان پیشنهاد صفحات مختلفی وجود دارد اما مطابق با جدول (۲) در این آزمایش با ۵۰۰ کاربر آموزش داده شده در اکثر خوشه‌های منتخب ۱۰ صفحه نشان داده شده ارجح‌تر هستند. اعتبارسنجی در سیستم‌های توصیه‌گر غالبا بر روی موارد مرتبط با دقت، فراخوانی[1] و پوشش[2] استفاده می‌گردد.



شکل ۴. میزان علاقه‌مندی کاربران به صفحات در پایگاه داده مورد مطالعه

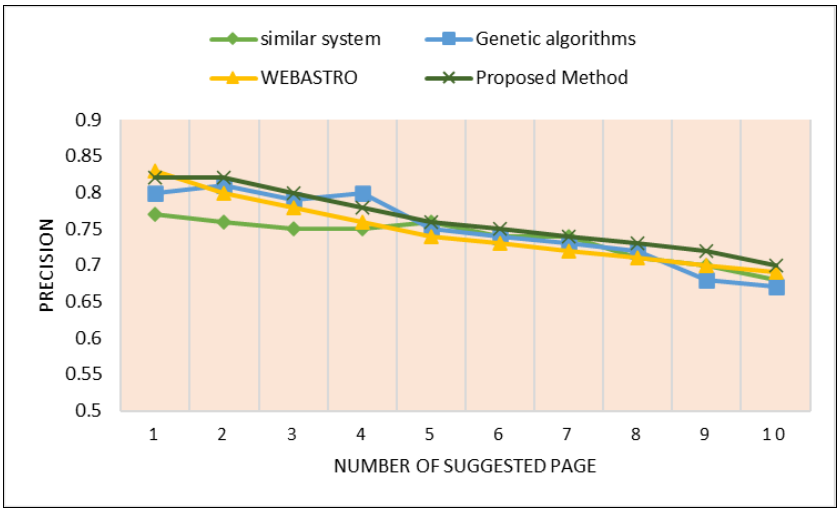
جدول ۲. ده صفحه پیشنهادی مهم با توجه به میزان علاقه‌مندی

نام صفحه	اولویت	نام صفحه	اولویت
lcc	6	ksc	1
apollo	7	mission-sts-71	2
apollo-15-info	8	tour	3
procurement	9	history	4
sts-gnnc	10	mission-sts-70	5

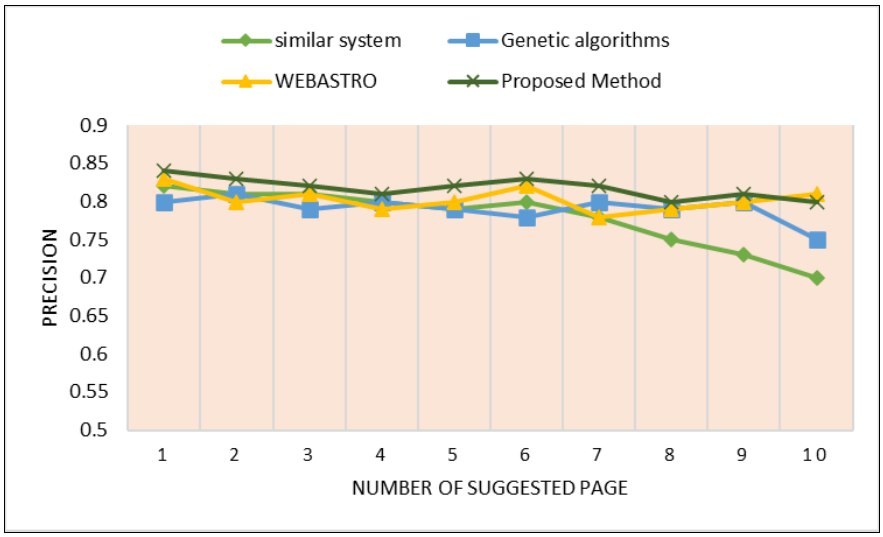
برای ارزیابی و مقایسه سیستم از سیستم مشابه در (Chidambaram, et al, 2018; Le, et al, 2024; Thwe, 2021) ارائه گردیده،

[1] recall
[2] coverage

استفاده شده است. شکل (۵) اعمال مراتب مختلف مدل بردار پشتیبان بر روی سیستم را نشان می‌دهد. همانطور که مشاهده می‌شود میزان دقت سیستم در مراتب بالاتر مدل تغییری نکرده است؛ به همین دلیل در این تحقیق برای کاهش میزان پیچیدگی سیستم، مدل بردار پشتیبان برای اعمال بر روی سیستم انتخاب شده است تا علاوه بر دارا بودن دقت مناسب، از افزایش پیچیدگی سیستم جلوگیری گردد. نتایج حاصل از ارزیابی پیاده‌سازی در ادامه نشان داده شده است.

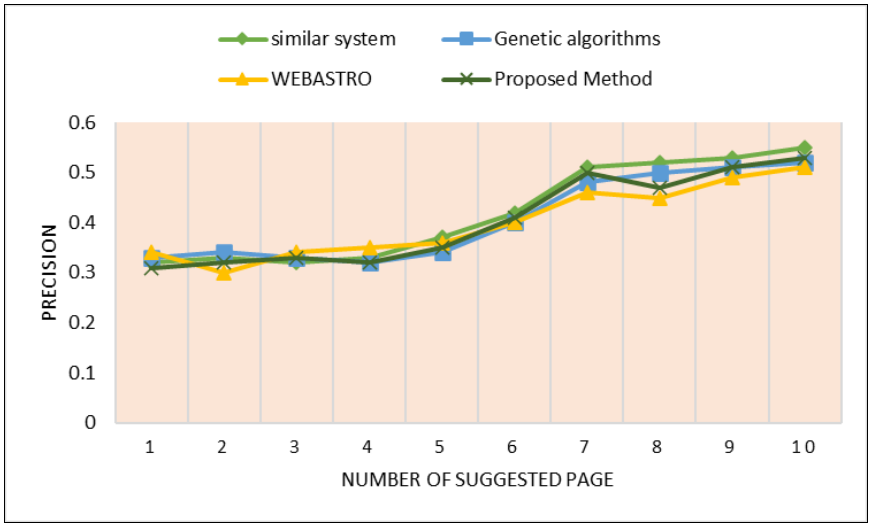


شکل ۵. ارزیابی دقت (Chidambaram, et al, 2018; Le, et al, 2024; Thwe, 2021)



شکل ۶. ارزیابی فراخوانی (Chidambaram, et al, 2018; Le, et al, 2024; Thwe, 2021)

با بیشتر شدن تعداد صفحات پیشنهادی دقت سیستم کاهش می‌یابد، یعنی تعداد صفحات پیشنهادی درست کمتری نسبت کل صفحات، پیشنهاد شده است. طبق نمودارهای به دست آمده و همانطور که انتظار می‌رفت، میزان دقت سیستم پیشنهادی نسبت به سیستم مشابه بهبود یافته است. این بهبود دقت همراه با کاهش زیاد حجم داده‌ها ورودی به سیستم می‌توان مزیت بسیار خوبی برای این سیستم نسبت سیستم مشابه باشد.



شکل ۷. ارزیابی پوشش (Chidambaram, et al, 2018; Le, et al, 2024; Thwe, 2021)

بحث و نتیجه‌گیری

روش پیشنهادی ارائه شده توانسته است حجم بسیار زیادی از اطلاعاتی را که میزان اهمیت آن‌ها پایین می‌باشد را حذف نماید و با این کار میزان زمان پاسخگویی سیستم را کاهش داد. زمان پاسخگویی سریع یکی از ارکان اصلی هر سیستمی می‌باشد که پایین بودن آن نشان از کارایی سیستم مذکور می‌باشد. استفاده از سیستم خوشه‌بندی بهینه شده یکی دیگر از مزایای این سیستم در کاهش میزان پیچیدگی آن می‌باشد. به طوری که با کوچک کردن ماتریس‌های به کار رفته در آن با حذف موارد غیر لازم میزان فضای مورد استفاده برای ذخیره‌سازی آنها را نیز کاهش می‌دهد. این سیستم با حذف اشتباهات کاربران در انتخاب پیوندها و استفاده از مدل پیش‌بینی کارآمد می‌تواند باعث افزایش میزان دقت کلی سیستم شود و به این ترتیب پیشنهادهای مناسب‌تری را به کاربران ارائه دهد. روش ارائه شده در این تحقیق بر اساس وب‌کاوی مبتنی بر کاربرد ارائه گردیده است که می‌تواند بدون نیاز به دخالت کاربران و به صورت کاملا صریح، نیازها و علاقه آنان را تشخیص داده و پیشنهادهای مناسب را به آن‌ها ارائه نماید. هدف از این تحقیق ارایه روشی جهت بهبود عملکرد موتورهای جستجو با استفاده از وب‌کاوی مبتنی بر کاربرد و بردار پشتیبان می‌باشد که نتایج به دست آمده از این تحقیق حاکی از کارا بودن این سیستم در کاربردی بودن آن در موارد می‌باشد. در حقیقت ارائه صفحات متناسب با نیازمندی کاربران به صورت خودکار و با دقت و کارایی بالا علاوه بر کاهش زمان پاسخ‌دهی و افزایش پوشش‌دهی نیازهای کاربران، سرویس‌دهنده نیز مشمول سوددهی واقع می‌شود، بعبارتی سرویس‌دهنده با ارائه محصولات و سرویس‌های خود در این امر نیز موفق می‌شود که با شناخت علایق کاربران بهترین تصمیم‌ها را در راستای پیشرفت و سوددهی خود انجام دهد.

پیشنهادهایی که برای کارهای آینده در این زمینه می‌باشند عبارتند از:

- یکی از موارد اطلاعاتی موجود در فایل‌های ثبت وب آدرس IPکاربران می‌باشد. با استفاده از این اطلاعات می‌توان از موقعیت جغرافیایی کاربران اطلاع حاصل کرد. از این اطلاعات می‌توان در سیستم پیشنهاد دهنده برای ارائه پشنهادهای مربوط به آن موقعیت جغرافیایی استفاده نمود.
- استفاده از سیستم ردیابی حرکات چشم یکی دیگر از روش‌هایی می‌باشد که می‌تواند در این گونه سیستم‌های پیشنهاد دهنده به کار آید. از آنجا که این سیستم‌ها به صورت صریح و بدون دخالت کاربر عمل می‌نمایند این تکنیک می‌تواند بسیار مفید واقع شده و اطلاعات بسیار مهمی را از علاقه کاربران در اختیار سیستم قرار دهد.

- با توجه به مدت زمان مشاهده صفحات توسط کاربران، می‌توان پی برد که مدت زمان بازدید از یک صفحه خاص توسط کاربران مختلف، متفاوت می‌باشد. به نظر می‌رسد با جداسازی این کاربران با توجه به مدت زمان بازدید آن صفحه و قرار دادن کاربران با مدت زمان بازدید مشابه در یک خوشه، می‌توان به دقت بالاتری از پیش‌بینی در سیستم پیشنهادی دست یافت.

منابع

کاظمی ،شهره ،(۱۳۹۳) " ترکیب مدل مارکوف به همراه روش خوشه بندی دوره های کاربر برای پیش بینی صفحه بعدی" ، همایش ملی مهندسی رایانه و مدیریت فناوری اطلاعات، تهران.

مقیمی، مهدی و همکاران،(۱۳۹۳) " تلفیق خوشه بندی و مدل مارکوف در یک چارچوب جدید برای پیشبینی صفحه بعدی انتخابی توسط کاربر"، نوزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، تهران.

Jaiganesh, M., Ramadoss, B., Vincent, A. and Mercy, S. (2020). Performance Evaluation of Cloud Services with Profit Optimization, Procedia Computer Science, vol. 54, pp. 24.

Robertson, A. M., & Willett, P. (2019), "An upperbound to the performance of rankedoutput searching: optimal weighting of query terms using a genetic algorithm". Journal of Documentation, Vol 4, pp.405–420.

Okabe, M., Seiji, Y., (2017), "Learning Filtering rulesets for ranking refinement in relevance Feedback", Knowledge-Based Systems 18 , pp117-124.

Chintan. R ,.Varnagar. N, Madhak. T, Kodinariya. M, NRathod. J., (2019), "Web Usage Mining: A Review on Process, Methods and Techniques", IEEE ,Information Communication and Embedded Systems (ICICES), International Conference , pp: 40 .

Priyanka S. Panchal, Prof. Urmi D. Agravat.,(2020), "Hybrid Technique for User's Web Page Access Prediction based on Markov Model", IEEE 4th ICCCNT, Computing, Communications and Networking Technologies, pp. 1 – 8.

Meera. N, and Sakina Banu. SH., (2015), "Predicting user's web navigation behavior using hybrid approach", Procedia Computer Science 45,pp .3-12.

Madhuri, B. C., Chandulal, A. J., Ramya, K., & Phanidra, M. (2019). "Analysis of users' web navigation behavior using grpa with variable length markov chains", International Journal of Data Mining & Knowledge Management Process, 1(2), pp .1-20.

Zhu. J, Hong. J, and G. Hughes. J., (2018), "Using Markov Chains for Link Prediction in Adaptive Web Sites", Soft-Ware, LNCS 2311, pp. 60–73.

Montgomery., L. A, Li. SH, Srinivasan. K, and C. Liechty. J., (2020), "Modeling Online Browsing and Path Analysis Using Clickstream Data", Marketing science, 23(4), pp.579.

Kalbhor, M., & Jain, K. (2019), "Fuzzy based hybrid approach for user request prediction using Markov model", In Computer, Communication and Control (IC4), vol 3, pp501.

Yang, Z., Wu, B., Zheng, K., Wang, X., & Lei, L., (2018), "A survey of collaborative filtering-based recommender systems for mobile internet applications". IEEE Access, Vol 4, pp.3273-3287.

Katarya, R., & Verma, O. P. (2017). "An effective collaborative movie recommender system with cuckoo search", Egyptian Informatics Journal, 18(2), pp.105-112.

Liu, H., Wu, J., Liu, T., Tao, D., Fu, Y. (2017). "Spectral Ensemble Clustering via Weighted K-Means: Theoretical and Practical Evidence", IEEE Transactions on Knowledge and Data Engineering, Vol 29, pp.1129 – 1143.

Chidambaram, S., Srinivasagan, K.G., (2018), "Performance evaluation of support vector machine classification approaches in data mining", Springer in Cluster Computing , Vol3, pp.1–8.

Le, T. D., Le-Dinh, T., & Uwizeyemungu, S. (2024). "Search engine optimization poisoning: A cybersecurity threat analysis and mitigation strategies for small and medium-sized enterprises", Technology in Society, 76, 102470.

Thwe, P. (2021). "Web page access prediction based on integrated approach", International Journal of Computer Science and Business Informatics, 12(1), pp.55-64.

A Method To Optimize Search Engine Results Using Clustering And Support Vector Machine (SVM) Algorithms

Abstract

Nowadays, the knowledge and information of the web is increasing, and the number of websites is also increasing day by day; As a result, there is a need to provide conditions in discussions such as e-commerce, search engines, targeted advertising, etc. based on the approach of beginner and professional users from different pages. But to be able to predict the needs and interests of different users and to suggest their favorite pages is associated with complications and it should be possible to do this with optimal methods and by using the movement patterns of users recorded in the web servers. did the job Various methods for predicting web pages have been presented, but the measures taken in this field have not yet reached a level that provides full user satisfaction, because each of the techniques available in this field has its advantages and disadvantages. Times and conditions are different, and for this reason, one technique cannot be determined to be better than other techniques in all aspects. Therefore, in order to achieve the desired quality, more research should be done so that the obtained results have high and acceptable accuracy and the way of doing the work is also a little complicated. In this research, a method to improve the performance of search engines using clustering and support vector has been presented, which consists of several different steps, including data pre-processing, creating a user registration profile, extracting the level of interest of the user, clustering and support vector. has been In addition to reducing the volume of information, the presentation of the proposed method has reduced the complexity of the system, and in the forecasting stage, by making changes in the support vector model, the accuracy of the system has also increased compared to the base model. The support vector model and the clustering method used in this system worked optimally to obtain acceptable results from the system. The evaluation criteria in this research are precision, recall and coverage. The overall accuracy of the system is equal to 83.5. The evaluation results of this system with the help of MATLAB tool with other similar tools show better and optimal performance.

Keywords web mining, clustering, user behavior patterns, support vector