

ارائه‌ی یک سیستم توصیه‌گر جدید با استفاده از فرایندهای مارکوف همه مراتب اصلاح شده

رقیه زارع^۱، علی هارون آبادی^۲

^۱ کارشناس ارشد نرم افزار، دانشگاه آزاد اسلامی واحد شبستر، rg.zare@gmail.com

^۲ استادیار، استادیار دانشگاه آزاد اسلامی واحد تهران مرکز، a.harounabadi@gmail.com

چکیده

کاوش رفتارهای کاربران در وب، به منظور پیش‌بینی رفتار بعدی آنها، به عنوان یکی از ابزارهای مهم در حوزه‌ی وب‌کاوی شناخته می‌شود. بهبود فرآیند پیش‌بینی می‌تواند موجب کاهش زمان دسترسی کاربران به اطلاعات مورد نیازشان در هنگام پیمایش صفحات وب گردد. مدل‌های مارکوف به طور گسترده برای مدل‌سازی و پیش‌بینی رفتار پیمایشی کاربر استفاده می‌شوند. این مدل براساس احتمال انتقال بین صفحات وبی که قبلاً در فایل‌های ثبت وب ذخیره شده‌اند، عمل می‌کند. در زمینه‌ی پیش‌بینی صفحات وب آنچه که از اهمیت اصلی برخوردار است، صحت پیش‌بینی می‌باشد. از این رو، راهکار ارائه شده در این مقاله صحت پیش‌بینی را نسبت به مدل مارکوف همه مراتب افزایش می‌دهد. افزون بر این، راه حل ارائه شده در مقایسه با قوانین انجمنی که یکی از روش‌های استفاده شده در وب‌کاوی جهت پیش‌بینی وب می‌باشند، از صحت پیش‌بینی بالاتری برخوردار است.

واژه‌های کلیدی

وب‌کاوی، مارکوف، خوشه‌بندی، مارکوف همه مراتب

۱- مقدمه

که بتواند محتمل‌ترین صفحه‌ای را که کاربر به عنوان صفحه بعدی خواهد دید، پیش‌بینی کند [3].

در زمینه‌ی پیش‌بینی صفحه‌ی وب با استفاده از مدل مارکوف، راهکارهایی ارائه شده است که یکی از آنها مدل مارکوف همه مراتب [5] می‌باشد. هدف مورد نظر در این مقاله، ارائه‌ی یک روش جدید و البته قدرتمند جهت بهبود صحت پیش‌بینی مدل مارکوف همه مراتب می‌باشد. همچنین از دیگر اهداف مقاله، ارائه‌ی یک معیار جدید برای وزندهی صفحات می‌باشد که از این معیار برای خوشه‌بندی نشست‌های کاربران استفاده شده است. در نتیجه، کاربرانی که علایق مشترکی دارند در یک گروه قرار گرفته و این امر، قدرت پیش‌بینی رفتار پیمایشی کاربران را افزایش می‌دهد.

۲- مدل مارکوف

رفتار یا سلیقه‌ی کاربر یک سایت بر اساس زنجیره‌ی صفحاتی که توسط وی پیمایش شده است، مدل می‌گردد. این مجموعه صفحات به عنوان یه دوره (w) در نظر گرفته شده و به صورت دنباله‌ای از صفحات که توسط کاربر بازدید شده‌اند، نمایش داده می‌شود.

وب‌کاوی به معنای بکارگیری تکنیک‌های داده‌کاوی برای کشف و استخراج خودکار اطلاعات از اسناد و سرویس‌های وب می‌باشد. بطور کلی وب‌کاوی را می‌توان داده‌کاوی بر روی داده‌های محتوا، ساختار و کاربرد وب (فایل‌های ثبت وقایع در سرویس‌دهنده‌های وب) قلمداد کرد. در صورتی که تکنیک‌های داده‌کاوی بر روی داده‌های کاربردی وب به منظور استخراج دانش مفید در رابطه با رفتار کاربران استفاده شود، وب‌کاوی را کاوش استفاده از وب می‌نامند [2].

پیش‌بینی صفحات بعدی مورد علاقه‌ی کاربر یکی از زمینه‌های تحقیقاتی کاوش استفاده از وب می‌باشد. در کاوش استفاده از وب از تکنیک‌های عمده‌ی داده‌کاوی از جمله کاوش قوانین انجمنی، خوشه‌بندی، طبقه‌بندی و استخراج الگوهای ترتیبی استفاده می‌شود [2]. در این بین، مدل مارکوف نیز یکی از متداول‌ترین روش‌ها می‌باشد که با توجه به رفتار گذشته‌ی کاربر، رفتار بعدی کاربر را پیش‌بینی می‌کند. مدل‌های مارکوف برای مطالعه و بررسی فرایندهای تصادفی استفاده می‌شوند - که برای مدل کردن پیش‌بینی رفتار پیمایشی کاربر بر روی وب سایت نیز خوب عمل کرده‌اند. ورودی این مسائل در حالت کلی، دنباله‌ای از صفحات وب است که به وسیله‌ی کاربر ملاقات شده‌اند و نتیجه‌ی آن ساختن مدل مارکوفی است

اگر دوره را $W_n = \langle P_1, P_2, P_3, \dots, P_n \rangle$ فرض کنیم، مسئله‌ی پیش‌بینی صفحه‌ی بعدی به این صورت تعریف می‌شود که با داشتن دوره- W بتوان صفحه‌ی P_{n+1} را پیش‌بینی کرد. مسئله‌ی پیش‌بینی صفحه‌ی بعدی می‌تواند توسط یک روش احتمالی به صورت زیر حل شود.

فرض کنید W_n دوره‌ی وب کاربر با طول n باشد و $P(P_i|W_n)$ احتمال اینکه صفحه‌ی P_i به عنوان صفحه‌ی بعد دیده خواهد شد. سپس صفحه‌ی P_{n+1} از فرمول زیر مشخص می‌شود:

$$P = \operatorname{argmax}_{p \in \rho} \{P(p_{n+1} = p | w)\} = \operatorname{argmax}_{p \in \rho} \{P(p_{n+1} = p | P_n, P_{n-1}, P_{n-2}, \dots, P_1)\} \quad (1)$$

که در آن ρ مجموعه‌ی کل صفحات موجود بر روی سایت است. در اصل با این روش احتمال تمام صفحات برای صفحه‌ی بعدی بودن محاسبه و سپس صفحه‌ای با بیشترین احتمال به عنوان پیش‌بینی انتخاب می‌شود. اما به دلیل اینکه محاسبه‌ی تمام این احتمالات شرطی به صورت دقیق غیرممکن است، از فرایند مارکوف برای پیش‌بینی صفحه‌ی بعدی استفاده می‌شود. با توجه به این فرایند، تنها آخرین k صفحه‌ی دیده شده توسط کاربر برای پیش‌بینی استفاده می‌شود که $k \ll n$ می‌باشد و تعداد صفحات (k) مشخص‌کننده‌ی مرتبه‌ی مدل مارکوف می‌باشد و فرمول مشخص کردن صفحه‌ی P_{n+1} به صورت زیر می‌باشد [4]:

$$P_{n+1} = \operatorname{argmax}_{p \in \rho} \{P(p_{n+1} = p | P_n, P_{n-1}, P_{n-2}, \dots, P_{n-(k-1)})\} \quad (2)$$

بنابراین در پیش‌بینی وب، منظور از مدل مارکوف مرتبه‌ی K ام، احتمال مشاهده‌ی صفحه‌ی k ام توسط کاربر می‌باشد که قبلاً $k-1$ صفحه را مشاهده کرده است. برای مثال در مارکوف مرتبه‌ی دوم، پیش‌بینی صفحه‌ی بعدی بر اساس دو صفحه‌ای که قبلاً مشاهده شده است صورت می‌گیرد. مزیت اصلی مدل مارکوف، کارایی و عملکرد آن بر حسب مدل ساخت و زمان پیش‌بینی است. به سادگی می‌توان نشان داد که زمان مورد نیاز برای ساخت مدل مارکوف با مرتبه‌ی k ام با اندازه‌ی مجموعه‌ی آموزش، به صورت خطی می‌باشد [3].

۳- کارهای گذشته

پژوهش‌های فراوانی در حوزه‌ی کاوش استفاده از وب و پیش‌بینی رفتار کاربر انجام شده است که در آنها فرایندهای مارکوف و خوشه‌بندی نیز به چشم می‌خورد. ضعف مدل مارکوف مرتبه‌ی پایین‌تر (مرتبه‌ی یک و دو)، عدم توانایی در پیش‌بینی دقیق صفحه‌ی بعدی و عدم پرداختن به گذشته-ی کاربر می‌باشد. مارکوف مرتبه‌ی بالاتر نیز دارای مشکلاتی از قبیل تعداد بالای حالت‌های آن و پوشش کم می‌باشد که گاه به خاطر پوشش کم، دقت کم‌تری نیز دارد. به منظور حل مشکل پوشش، می‌توان از مارکوف همه مراتب استفاده کرد.

در مقاله‌ی [5] برای رفع مشکل دقت و پوشش پایین‌تر مارکوف مراتب پایین، از مارکوف همه مراتب استفاده شده است. شیوه‌ی کار این مدل بدین صورت است که برای انجام عمل پیش‌بینی به ازای هر نمونه، از بزرگترین مرتبه‌ی مدل مارکوفی که نمونه را پوشش می‌دهد، استفاده می-

کند. برای مثال اگر این مدل شامل سه مرتبه از مدل مارکوف باشد، نمونه-ی داده شده ابتدا در صورت امکان با مرتبه‌ی سه پیش‌بینی می‌شود و در صورتی که توسط مرتبه‌ی سه پوشش داده نشود، این روند برای مرتبه‌ی دو و یک هم تکرار می‌شود.

در روش ارائه شده در این مقاله در فاز پیش‌بینی، از هر مرتبه‌ی مارکوف زمانی استفاده می‌شود که امکان پیش‌بینی با آن مرتبه وجود داشته باشد. یعنی به طور همزمان از اطلاعات دو مرتبه‌ی مدل مارکوف استفاده نمی‌شود.

در پژوهش حاضر، از مقادیر پیش‌بینی شده‌ی هر مرتبه از مدل مارکوف استفاده شده و روشی بهینه ارائه خواهد شد که هم مدل مارکوف همه مراتب را بکار گرفته و هم نسبت به مدل مارکوف همه مراتب دارای صحت پیش‌بینی بالاتری می‌باشد.

یکی دیگر از معایب مارکوف همه مراتب، پیچیدگی فضایی یعنی تعداد زیاد حالت‌های آن می‌باشد که در این حوزه، روش‌های مختلفی به منظور کاهش تعداد حالت‌ها ارائه شده است.

در زمینه‌ی پیش‌بینی وب آنچه که مهم است، صحت پیش‌بینی به همراه کاهش زمان مورد نیاز برای پیش‌بینی فاز برخط می‌باشد. در [1] از مدل خوشه‌بندی کامیانه استفاده شده و توسط یک چارچوب جدید قابل گسترش، به پیش‌بینی صفحه‌ی بعدی کاربر پرداخته شده است. آنها توانسته‌اند به میزان قابل قبولی صحت کار را افزایش دهند. یکی از اهداف نویسندگان، ایجاد یک روش هوشمند برای انتخاب بهترین مرتبه‌ی مدل مارکوف می‌باشد که باعث کاهش پیچیدگی پیش‌بینی در فاز برخط می‌شود. بدین منظور روشی مبتنی بر برچسب ارائه شده است که علاوه بر کاهش پیچیدگی فرایند، صحت کار را به میزان قابل قبولی افزایش می‌دهد. آزمایشات انجام شده نشان می‌دهد که روش پیشنهادی در این پژوهش نسبت به روش ارائه شده در [1] از صحت پیش‌بینی بالاتری برخوردار است.

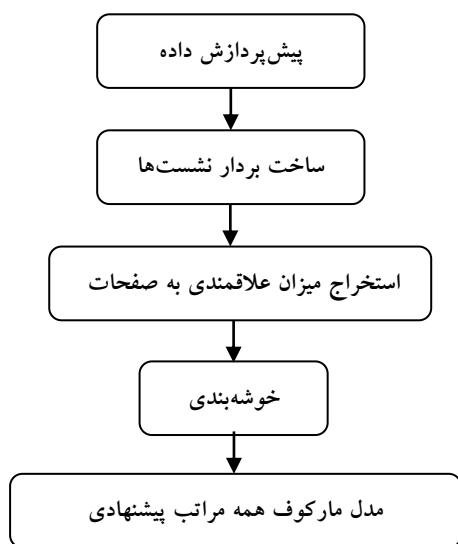
در [6] از دو نوع داده جهت پیش‌بینی استفاده شده است که این داده-ها عبارتند از داده‌های فایل ثبت و دیگری داده‌های محتوا که شامل کلمات کلیدی هر صفحه می‌باشند. روش پیشنهادی در سه ماژول پیاده‌سازی شده است. ماژول اول نشست‌های کاربران را در قالب بردار نشست از وب سایت جمع‌آوری کرده و سپس آنها را خوشه‌بندی می‌کند. ماژول دوم به جهت بهبود احتمال پیش‌بینی وب، کلمات کلیدی هر صفحه‌ی وب را جمع‌آوری می‌کند. در این ماژول هنگامی که نشست جدیدی وارد شود، ابتدا کلمات کلیدی آن جمع‌آوری می‌شوند. در ماژول سوم، از مدل مارکوف همه مراتب جهت پیش‌بینی استفاده می‌شود. در این مرحله صفحاتی که بوسیله‌ی مدل مارکوف پیش‌بینی می‌شوند، از طریق صفحه‌ای که بیشترین کلمات مورد علاقه‌ی کاربر را دارد فیلتر شده و بهترین صفحه به کاربر پیشنهاد داده می‌شود.

در مقاله‌ی [7] از چندین ویژگی پیمایشی جهت پیش‌بینی صفحه‌ی بعدی استفاده می‌شود. شباهت صفحه، اندازه‌ی صفحه، زمان دسترسی

تواند به طور قطعی عمل کند و اطلاعات زیادی در مورد گذشته‌ی کاربر مورد نیاز است. مدل پیشنهادی، ابتدا از روش کاهش حالت مدل مارکوف که در [4] ارائه شده است جهت حذف برخی حالات استفاده کرده و پس از آن مارکوف مرتبه‌ی دو را به کار می‌گیرد. آزمایشات انجام شده نشان می‌دهد که روش پیشنهادی در این پژوهش نسبت به روش ارائه شده در [11] از صحت پیش‌بینی بالاتری برخوردار است

۴- روش پیشنهادی

روش پیشنهادی از دو بخش برون خط و برخط تشکیل شده است. در مرحله‌ی برون خط ابتدا عملیات پیش‌پردازش بر روی فایل‌های ثبت انجام می‌شود تا کلیه‌ی نشست‌های کاربران بدست آید. سپس نشست‌های کاربران با استفاده از الگوریتم کامیانه خوشه‌بندی شده و مدل مارکوف پیشنهادی بر روی هر خوشه اعمال می‌شود. مرحله‌ی نهایی، ورود نشست تست جدید و انجام عمل پیش‌بینی برای آن نشست با استفاده از مدل مارکوف و اطلاعات خوشه‌ی مربوطه می‌باشد. توضیحات بیشتر پیرامون مراحل انجام این کار در شکل ۱ آمده است.



شکل ۱: روش پیشنهادی

۴-۱- پیش‌پردازش داده

ثبت‌های سرور به عنوان یک منبع اطلاعاتی بزرگ و اصلی برای استخراج دانش در حوزه وب‌کاوی در نظر گرفته می‌شوند. داده‌های ذخیره شده در ثبت‌های سرور نشان‌دهنده‌ی دسترسی چندین کاربر به یک وب-سایت می‌باشند. پیش‌پردازش داده‌ها شامل فیلتر کردن داده‌ها، شناسایی کاربران و شناسایی نشست‌ها می‌باشد. در فایل‌های ثبت، همه‌ی داده‌های ثبت برای کاوش استفاده از وب مناسب نیستند و تنها باید داده‌هایی را نگه داریم که اطلاعات مناسبی دارند. بنابراین برای حذف کردن داده‌های نامناسب از فایل‌های ثبت، عمل پاکسازی روی آنها انجام می‌شود. از آنجا که عمل پیش‌پردازش وابسته به دامنه است، نسبت به دامنه‌ی وب‌سایت‌ها لازم است که بین داده‌های مرتبط و غیرمرتبط تمایز قائل

صفحه، مدت زمان صفحه و فرکانس صفحه یا انتقال از جمله‌ی این موارد می‌باشند. شیوه‌ی کار در این مقاله بدین صورت است که ابتدا مدل مارکوف را بر روی نشست وب اعمال می‌کند و اگر نتایج عمل پیش‌بینی مبهم باشد یا به عبارت دیگر چند صفحه با احتمالات یکسان پیش‌بینی شده باشد، برای افزایش دقت پیشنهاد صفحه از الگوریتم رتبه‌بندی صفحه بر مبنای محبوبیت و تشابه (PSPR) استفاده می‌کند.

در مقاله‌ی [8] یک مدل مارکوف اصلاح شده‌ی جدید پیشنهاد شده است که در آن همه‌ی نشست‌هایی را که شامل صفحات مشابه ولی با ترتیب‌های مختلف هستند، یکسان در نظر می‌گیرد. انگیزه‌ی آنها برای این کار آن است که یک عمل روی وب می‌تواند با استفاده از مسیرهای مختلف صرفنظر از ترتیبی که کاربران انتخاب می‌کنند، انجام شود. نویسندگان مقاله همچنین اندازه‌ی مدل پیش‌بینی را با حذف نشست‌هایی که صفحات در آنها تکرار شده‌اند، کاهش می‌دهند. این عمل با وجود کاهش اندازه‌ی مدل، هیچ تأثیر سوئی بر دقت مدل ندارد. علاوه بر این، یک چارچوب پیش‌بینی دو لایه جهت بهبود زمان پیش‌بینی نیز پیشنهاد شده است.

مقاله‌ی شماره‌ی [9] از مدل مارکوف تودرتوی پویا برای مدل کردن نشست‌های رفتاری کاربر استفاده کرده است. در این مدل تمرکز اصلی بر روی پوشش و پیچیدگی مدل است. مشکل مدل مارکوف سنتی این است که شامل اطلاعاتی درباره‌ی تنها یک مرتبه‌ی خاص است و این مسأله، پوشش و دقت پیش‌بینی را کاهش می‌دهد. در مدل پیشنهادی، مدل مارکوف مراتب بالاتر در داخل مدل مارکوف مرتبه‌ی پایین‌تر قرار داده می‌شود. مدل پیشنهاد شده از ساختار لیست پیوندی به جای ماتریس انتقال مارکوف سنتی استفاده می‌کند.

در [10] یک مدل پیش‌بینی ترکیبی مؤثر به نام مدل پیش‌بینی دو مرحله‌ای ارائه شده است که از رفتار کاربران وب سایت بهره می‌گیرد. این مدل با دسته‌بندی صفحات در فایل ثبت سرویس‌دهنده بر اساس استفاده‌ی کاربران، تعداد صفحات کاندید برای انجام عمل بررسی را کاهش می‌دهد و در نتیجه باعث افزایش سرعت بارگزاری می‌شود. این مدل برای پیش‌بینی دسته کاربر جاری، از مدل مارکوف مرتبه‌ی دوم و ماتریس حالت ایستای انتقال استفاده نموده و سپس جهت پیش‌بینی صفحات مورد علاقه‌ی کاربر از قانون بیز بهره گرفته است. نتایج مقاله نشان می‌دهد که این مدل به طور چشمگیری زمان اجرا و پیش‌بینی را نسبت به مدل مارکوف مرتبه‌ی اول بدون دسته‌بندی صفحات بهبود می‌دهد؛ به ویژه هنگامی که تعداد صفحات وب در حال افزایش باشد.

روش پیشنهادی [11] ترکیب سه مدل خوشه‌بندی و مارکوف مرتبه‌ی دوم و قوانین انجمنی می‌باشد. به منظور کاهش تعداد تراکنش‌های استفاده شده، در ابتدا نشست‌های کاربران بر اساس معیارهای انتخاب ویژگی در داخل گروه‌هایی تقسیم می‌شوند. سپس گروه‌های نشست‌های وب با استفاده از الگوریتم خوشه‌بندی کامیانه و معیار فاصله‌ی کسینوسی خوشه‌بندی می‌شوند و عمل پیش‌بینی با استفاده از مارکوف انجام می‌شود. قوانین انجمنی زمانی مورد استفاده قرار می‌گیرند که مدل مارکوف نمی‌

شویم. در غیر اینصورت ممکن است این امر سبب از دست رفتن اطلاعات باارزش شود. رکوردهای ایجاد شده توسط خزشگرها و اسپایدرها در طول عمل فیلتر کردن حذف می‌شوند؛ چرا که به عنوان داده‌های کاربردی محسوب نمی‌شوند. همچنین در این مرحله تمامی درخواست‌های عکس، فیلم، صوت و انواع پسوندهای فایل‌ها بجز فایل‌های صفحات وب فیلتر می‌شود [2].

پس از پاکسازی، شناسایی کاربران مهم‌ترین مرحله در پیش‌پردازش داده‌هاست. چندین راه برای شناسایی کاربران وجود دارد که ساده‌ترین آن‌ها، استفاده از آدرس IP می‌باشد. روش دیگر، استفاده از کوکی‌ها و یا ترکیب آدرس IP با فیلد ارجاع است. همچنین با استفاده از نوع مرورگر و سیستم عاملی که کاربر از آن استفاده می‌کند به همراه IP، می‌توان کاربران مختلف را شناسایی کرد. در این سیستم پیشنهادی، از آدرس IP برای شناسایی کاربر استفاده می‌شود.

پس از شناسایی کاربران، بایستی نشست‌های کاربران شناسایی شود. یک نشست کاربر، مجموعه‌ای از صفحات ملاقات شده توسط کاربر در طول یک بازدید خاص او از وب سایت است [2]. در حالت کلی دو روش برای تشخیص نشست وجود دارد: روش مبتنی بر زمان و روش مبتنی بر پیمايش. در مقاله‌ی ارائه‌شده، از روش مبتنی بر زمان استفاده شده است. بعد از شناسایی نشست‌ها، به دو دسته‌ی "مجموعه نشست‌های آموزش" و "مجموعه نشست‌های تست" تقسیم می‌گردند. از مجموعه نشست‌های آموزش برای پیاده‌سازی ایده‌ی پیشنهادی و از مجموعه نشست‌های تست جهت ارزیابی سیستم استفاده خواهد شد.

۴-۲- ساخت بردار نشست‌ها

فرض کنید P مجموعه‌ی صفحات قابل دسترسی توسط کاربران یک سایت باشد به طوری که $P = \{p_1, p_2, \dots, p_m\}$ که هر صفحه با URL منحصر به فرد موجود است. همچنین $S = \{s_1, s_2, \dots, s_n\}$ مجموعه‌ی نشست‌های دسترسی کاربران است که در آن هر $s_i \in S$ زیرمجموعه‌ای از P است. هر نشست s_i با برداری m بعدی به صورت

$$s_i = \{w(p_1, s_i), w(p_2, s_i), \dots, w(p_m, s_i)\}$$

نمایش داده می‌شود که $w(p_j, s_i)$ وزن تعیین‌شده برای j امین صفحه‌ی وب بازدید شده در نشست s_i است [12].

۴-۳- استخراج میزان علاقمندی به صفحات

برای وزن‌دهی به صفحات باید میزان علاقه‌ی کاربر به صفحه مشخص شود. برای این منظور از "فرکانس صفحه" و "مدت زمان مشاهده‌ی صفحه" و "تاریخ مشاهده‌ی صفحه" استفاده شده است.

منظور از فرکانس صفحه، تعداد دفعات مشاهده‌ی یک صفحه‌ی وب بوده و مدت زمان مشاهده‌ی صفحه، برابر با مدت زمان صرف شده بر روی یک صفحه است. بدیهی است در یک نشست، ممکن است کاربر به یک صفحه چندین بار مراجعه کند که هر چه تعداد ارجاع‌ها به یک صفحه در یک نشست بیشتر باشد، آن صفحه در نشست مذکور نسبت به سایر

صفحات نشست از اهمیت بیشتری برخوردار است. همچنین مدت زمان مشاهده‌ی یک صفحه توسط کاربر، میزان اهمیت صفحه را نشان می‌دهد. زیرا اگر صفحه‌ای برای کاربر جذاب نباشد، صفحه را رد کرده و به صفحه‌ی دیگری مراجعه می‌کند. در حالی که اگر صفحه مورد علاقه‌ی کاربر باشد، زمان قابل ملاحظه‌ای را صرف مشاهده‌ی آن خواهد کرد. البته باید در نظر داشت که اگر طول یک صفحه‌ی وب کم باشد، زمان مشاهده‌ی آن صفحه نیز به تناسب آن کمتر می‌شود. لذا زمان مشاهده‌ی صفحه متناسب با اندازه‌ی صفحه است و در محاسبه‌ی اهمیت صفحه، این تناسب باید در نظر گرفته شود [12].

با توجه به اینکه صفحات وبی که اخیراً مشاهده شده‌اند نسبت به صفحاتی که قبلاً مرور شده‌اند علائق جدید کاربر را بهتر نشان می‌دهند، به همین دلیل از معیار تاریخ برای وزن‌دهی به صفحات استفاده شده و صفحات بر اساس تاریخ رؤیت، به صورت خطی افزایش وزن داده می‌شوند. صفحات مربوط به آخرین نشست‌های کاربر ضریب بالاتری نسبت به نشست‌های پیشین دارند.

بدین ترتیب با توجه به توضیحات ذکر شده "مدت زمان مشاهده‌ی صفحه"، "فرکانس صفحه" و "تاریخ رؤیت صفحه" سه معیار اساسی در تعیین اهمیت و میزان علاقه‌ی کاربر به صفحه می‌باشند. پارامتر $Duration$ بیانگر مدت زمان مشاهده‌ی صفحه و $Size$ اندازه‌ی صفحه برحسب بایت، $Number of visit$ تعداد دفعات تکرار صفحه در یک نشست می‌باشد.

برای محاسبه‌ی معیارهای ذکر شده، از فرمول‌های زیر استفاده می‌شود:

$$frequency(page) = \frac{number\ of\ visits\ (page)}{\sum_{page \in visited\ page} number\ of\ visits\ (page)} \quad (3)$$

$$Duration(page) = \frac{total\ duration(page)}{Max_{page \in visited\ page} \left(\frac{total\ duration(page)}{size(page)} \right)} \quad (4)$$

از آنجا که نسبت اهمیت هر سه پارامتر برابر می‌باشد، از میانگین هارمونیک این سه معیار برای وزن‌دهی صفحات استفاده گردید:

$$Interest(page) = \frac{3 \times frequency(page) \times Duration(page) \times date(page)}{frequency(page) + Duration(page) + date(page)} \quad (5)$$

۴-۴- خوشه‌بندی

در وب‌کاوی کاربردی از روش‌های مختلفی جهت گروه‌بندی کاربرانی که رفتارهای مشابهی دارند استفاده شده است. به طور مثال در [13] از شبکه‌ی عصبی، در [14] از روش جفت نزدیک‌ترین همسایه، در [15] از الگوریتم خوشه‌بندی تقسیم گراف و در [11][16] از الگوریتم کامیانه برای خوشه‌بندی نشست‌های کاربران استفاده شده است.

در این پژوهش بعد از بدست آوردن میزان علاقه‌ی هر کاربر، از خوشه‌بندی کامیانه جهت گروه‌بندی کاربران بر اساس علاقه‌ی آنها به صفحات

استفاده شده است. در این الگوریتم، نشست‌ها بر اساس اندازه‌گیری فاصله-ی کسینوسی گروه‌بندی می‌شوند.

۴-۵- مدل مارکوف پیشنهادی

به منظور افزایش سرعت و قدرت در روند کار، به جای ذخیره‌ی یک ماتریس بزرگ در حافظه، از جداول درهم‌سازی استفاده شد. به این صورت که ابتدا تمام نشست‌های مرحله‌ی آموزش با استفاده از پنجره‌ی لغزان یک الی چهار درهم‌سازی شده و نشست‌های بخش‌شده با این پنجره به همراه تعداد تکرار آنها و صفحه‌ی بعدی نمایش داده شده در آن، در جدول درهم-سازی ذخیره می‌شوند. در واقع جدول درهم‌سازی نشان می‌دهد که بعد از هر دنباله صفحات مشاهده‌شده توسط کاربر، صفحه‌ی بعدی رویت شده چه صفحه‌ای است و این صفحه چند بار بعد از این دنباله توسط کاربران مختلف مشاهده شده است.

بدین ترتیب بعد از عمل خوشه‌بندی، مدل مارکوف از مرتبه‌ی یک تا مرتبه‌ی چهار بر روی هر خوشه با استفاده از تکنیک درهم‌سازی اعمال می‌شود.

پس از مرحله‌ی آموزش، در فاز برخط که مربوط به مرحله‌ی تست می‌باشد، برای انجام عمل پیش‌بینی بعد از اینکه نشست جاری کاربر وارد شد، ابتدا تشخیص داده می‌شود که نشست جاری به کدام خوشه تعلق دارد. به این ترتیب نزدیک‌ترین خوشه به نشست کاربر تعیین می‌شود.

در سیستم پیشنهادی، از مدل مارکوف مرتبه‌ی یک تا مرتبه‌ی چهار استفاده می‌شود. بدین‌صورت که برای انجام عمل پیش‌بینی، ابتدا مدل مارکوف مرتبه‌ی یک مورد جستجو قرار می‌گیرد اگر دنباله‌ی مورد نظر یافت شود، n صفحه‌ای که دارای احتمال بیشتر به عنوان صفحه‌ی بعدی می‌باشند انتخاب شده و اطلاعات آنها ذخیره می‌شود. پس از بررسی‌ها و آزمایشات به عمل آمده، تعداد ۲۰ مناسب شناخته شد. سپس مدل مارکوف مرتبه‌ی دو بررسی شده و ۲۰ صفحه‌ای که دارای احتمال بیشتر می‌باشند، انتخاب می‌شوند. این عملیات تا مرتبه‌ی چهار انجام می‌شود. سپس مجموع تمامی احتمالات مراتب یک، دو و سه و مرتبه‌ی چهار مدل مارکوف محاسبه می‌گردد. این کار باعث می‌شود که احتمالات صفحات مشابه در هر مرتبه پررنگ‌تر شود. در نهایت صفحه‌ای که دارای بیشترین احتمال است، به عنوان صفحه‌ی بعدی پیشنهاد داده می‌شود.

در حین پیاده‌سازی به ازای هر مرتبه مدل مارکوف، تمام صفحات یکتای سایت در نظر گرفته شد. ولی نتایج نشان داد که تنها ۲۰ صفحه‌ی اول در پیش‌بینی تأثیر دارند و در نظر گرفتن تمام صفحات یکتای سایت جهت پیش‌بینی نه تنها پیچیدگی کار را بالاتر برده و تأثیری در صحت کار ندارد، بلکه این کار زمان پیش‌بینی را نیز افزایش می‌دهد.

۵- ارزیابی و آزمون

در مقاله‌ی حاضر، آزمایشات بر روی مجموعه داده‌ی NASA صورت پذیرفت. جدول شماره‌ی یک، مشخصات این مجموعه داده را نمایش می‌دهد.

جدول ۱: مشخصات مجموعه داده‌ی استفاده شده

	NASA
Total requests	3,461,612
Total sessions	73750
Num of pages	876
Dataset date	1-30/07/1995

برای محاسبه‌ی میزان صحت کار، از روش مورد استفاده در [8] و [16] بهره گرفته شده است. قبل از بررسی صحت باید به بررسی چند مفهوم پرداخت. در رویکرد حاضر، از مفهومی به نام رتبه استفاده شده است. منظور از Rank آن است که به جای یک صفحه، i صفحه‌ی بعدی کاربر پیش‌بینی شده و در صورتی که یکی از این i صفحه‌ی پیش‌بینی‌شده، صفحه‌ی واقعی بعدی کاربر باشد، آنگاه گفته می‌شود که پیش‌بینی درست انجام شده است.

Hit: در صورتی که پیش‌بینی صحیح باشد (یا در صورتی که از رتبه استفاده شده باشد، پیش‌بینی انجام شده در لیست پیش‌بینی‌ها قرار داشته باشد)، به Hit یک عدد اضافه می‌شود.

Miss: در حالتی که پیش‌بینی اشتباه باشد (یا در لیست پیش‌بینی موجود نباشد).

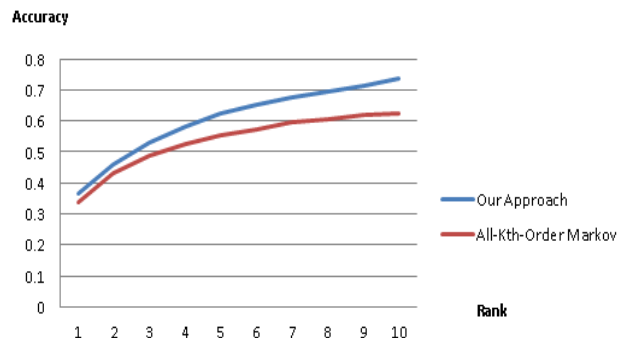
Match: در پژوهش حاضر، از جدول درهم‌سازی برای ذخیره‌سازی رفتار قبلی کاربران استفاده شده است که سرعت فرایند پیش‌بینی را بسیار افزایش می‌دهد. در صورتی که Match یک عدد اضافه می‌شود که بدون در نظر گرفتن پیش‌بینی صحیح یا غلط صفحه‌ی بعدی، جدول درهم‌سازی به عنوان خروجی، حداقل یک صفحه به کاربر پیشنهاد دهد.

برای تشخیص صحت پیش‌بینی در هر نشست، آخرین صفحه نگه داشته شده و برای بقیه صفحات، شروع به انجام عمل پیش‌بینی می‌شود. در انتها صحت پیش‌بینی برابر با نسبت پیش‌بینی‌های صحیح بر کل تعداد Match‌ها خواهد بود.

شکل ۲، صحت روش پیشنهادی را در مقایسه با مدل مارکوف همه مراتب نشان می‌دهند. نتایج آزمایش بیانگر آن است که الگوریتم پیشنهادی نسبت به این روش دارای صحت بالاتری است.

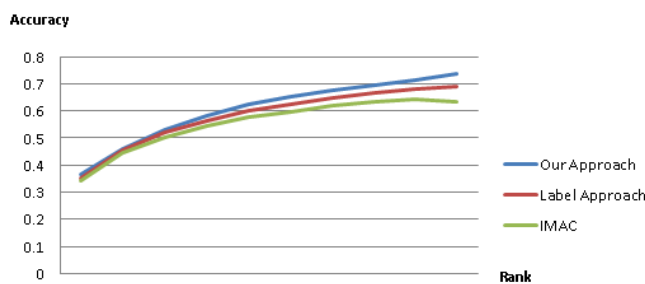
مراجع

- [۱] مقیمی، مهدی و همکاران "تلفیق خوشه‌بندی و مدل مارکوف در یک چارچوب جدید برای پیش‌بینی صفحه بعدی انتخابی توسط کاربر"، نوزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، دانشکده مهندسی کامپیوتر دانشگاه شهید بهشتی تهران، اسفندماه ۱۳۹۲.
- [2] R. Shukla, S. Silakari, and P.K Chande, "Web personalization systems and web usage mining: a review", International Journal of Computer Applications, vol. 72, pp. 6-13, June 2013.
- [3] S. Kuma, G. Abhinav, "A survey on markov model", MIT, International Journal of Computer Science & Information Technology, MIT, vol. 4, pp.29-33, Jan 2014.
- [4] M. Deshpande, G. Karypis, "Selective markov models for predicting web page accesses", ACM Trans. on Internet Technology, Vol. 4, pp.163-184, May 2004.
- [5] J. Pitkow, P. Pirolli, "Mining longest repeating subsequence to predict World Wide Web surfing", In 2nd USENIX Symposium on Internet Technologies and Systems, Boulder, CO, Octobr 1999
- [6] S. Vishwakarma, Lade, S. Raghavendr and M. kumar, "web user prediction by: integrating markov model with different feature", International Journal of Eneeneering Research an Science & Technology, Vol. 2, November 2013.
- [7] P. Thwe, "Proposed approach for web page access prediction Using Popularity And Similarity Based Page Rank algorithm", international journal of scientific & technology research, Vol 2, March 2013.
- [8] A. Mamoun, I. Khalil, "Prediction of user's web-browsing behavior: Application of markov model", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 42, Vol 42, pp. 1131-11, August 2012.
- [9] B. Nigam, S. Jain, "Generating a new model for predicting the next accessed web page in web usage mining", Third International Conference on Emerging Trends in Engineering and Technology, 2010 3rd International Conference on, IEEE, pp. 485 - 490, 2011.
- [10] C. Lee, Y. Lo and Y. Fu, "A novel prediction model based on hierarchical characteristic of web site", Expert Systems with Applications, Vol. 38, pp. 3422-3435, 2011
- [11] F. Khalil, j. Li, Wang, H., "An integrated model for next page access prediction. International", Journal of Knowledge and Web Intelligence, vol. 1, pp. 48-80, 2009.
- [12] H. Liu, V. Keşelj, "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. Data Knowl. Eng", Vol 61, pp. 304-330, 2007
- [13] K. Santos, V. Phoha and S. Balagani, "Adaptive Neural Network Clustering of Web Users", IEEE Computer, Vol 37, pp. 34-40, 2004
- [14] A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction", International Journal of Computer Applications, Vol. 8, no.11, October 2010
- [15] S. Gunduz, M. Ozsu, "A web page prediction model based on click-stream tree representation of user behavior", Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 24-27, 2003.
- [16] G. Poornalatha, S. Prakash, "Web page prediction by clustering and integrated distance", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Vol. 44, pp. 1349-1354, 2012.



شکل ۲: مقایسه‌ی صحت روش پیشنهادی با All-Kth-order Markov

شکل ۳، صحت روش پیشنهادی را در مقایسه با روش‌های ارائه شده‌ی [۱۱] و [1] نشان می‌دهد. نتایج آزمایش بیانگر آن است که الگوریتم پیشنهادی نسبت به هر دو روش دارای صحت بالاتری است.



شکل ۳: مقایسه‌ی صحت روش پیشنهادی با [11] و [1]

۶- نتیجه‌گیری

در این مقاله به بررسی مدل‌های مارکوف جهت پیش‌بینی وب پرداخته شد. سپس چارچوب پیشنهادی با جدیدترین روش‌های موجود در مقالات ارائه شده در جامعه‌ی علمی مورد مقایسه قرار گرفت. نتایج نشان داد که روش پیشنهادی نسبت به روش‌های بررسی شده دارای صحت بالاتری می‌باشد.

عمل خوشه‌بندی نشست‌ها بر حسب سه معیار که نشان دهنده‌ی علاقه‌ی کاربران به صفحات سایت می‌باشد، صورت پذیرفت. لازم به ذکر است که آزمایشات صورت گرفته نشان می‌دهد که انجام عمل پیش‌بینی بدون خوشه‌بندی باعث کاهش صحت پیش‌بینی می‌شود و همین امر نشان دهنده‌ی اهمیت خوشه‌بندی در پیش‌بینی می‌باشد.

آنچه که در پژوهش‌های پیش رو می‌تواند مورد نظر قرار گیرد، استفاده از داده‌های ساختار یا محتوا علاوه بر داده‌های کاربردی و تلفیق آن با مدل مارکوف در انجام عمل پیش‌بینی می‌باشد. همچنین می‌توان جهت افزایش صحت پیش‌بینی از سایر روش‌های خوشه‌بندی به‌جز کامیانه استفاده کرد. بعلاوه، بدست آوردن بهینه‌ی بهترین خوشه و در کنار آن، هدایت تست‌ها به بهترین خوشه‌ها به کمک روش‌های دیگری مانند شبکه‌ی عصبی به عنوان هدف آتی در توسعه‌ی این مقاله مورد توجه قرار خواهد گرفت.