مدل ریاضی تحلیل جریان کلیک برای پیش‌بینی رفتار مشتریان اینترنتی

محمدمهدی سپهری و فؤاد مهدوی‌پژوه

چکیده: تحلیل جریان کلیک ابزار مفیدی برای پیش‌بینی مسیر حرکت یک مشتری خاص در یک وب سایت است که کاربرد فراوانی در زمینه‌های تجارت الکترونیکی، بازاریابی الکترونیکی و مدیریت ارتباط با مشتری دارد. رویکرد جدید مقاله به‌دست آوردن محتمل‌ترین مسیر حرکت یک کاربر در یک وب سایت با استفاده از مدل‌های مارکوفی است که در قالب یک مدل برنامه‌ریزی صفر و یک حاصل شده است. مدل برنامه‌ریزی صفر و یک ارائه‌شده حالت خاصی از مدل معروف مسئله پیله‌ور (فروشنده دوره‌گرد) گردآورنده جایزه می‌باشد که خود یک مدل NP - hard بوده و تعداد محدودیت‌های حذف زیر تور آن با افزایش فضای مسئله به‌طور انفجارآمیزی افزایش می‌یابد. برای حل مدل طرح‌شده الگوریتمی جامع و کارا ارائه گردیده است. برای انجام جنبه‌های محاسباتی و پیاده‌سازی مدل پیشنهادی، داده‌های برگرفته از لاگ فایل‌های سرور یک وب سایت دانشگاهی برای ۲۰ کاربر مختلف مورد استفاده قرار گرفت. مقایسه جواب‌های حاصل با جواب‌های به‌دست آمده از الگوریتم جیودیچی نشان می‌دهد مدل پیشنهادی جواب‌های بسیار دقیق‌تر و بهتری نسبت به الگوریتم جیودیچی ارائه می‌دهد.

کلید واژه: **برنامه‌ریزی ریاضی، تحلیل جریان کلیک، مدل فروشنده دوره‌گرد گردآورنده جایزه، مدل زنجیره مارکوف.**

۱- مقدمه

روش‌ها و تکنولوژی‌های به‌کار رفته در دنیای امروز هر یک به‌طور اجتناب‌ناپذیری به سمت تولید حجم انبوهی از داده‌ها پیش می‌روند. در حال حاضر وب سایت‌های اینترنتی بزرگ‌ترین منبع تولید داده‌ها در دنیا هستند که در آنها این داده‌ها به اشکال مختلفی نظیر متن، عکس و سایر فرمت‌های صوتی و تصویری تولید می‌شوند. با توجه به محدودیت توانایی‌های انسان، حتی دیدن این حجم از داده‌ها هم برای بشر امکان‌پذیر نمی‌باشد. از این رو برای درک و استفاده مؤثر از این داده‌ها، نیازمند به‌کارگیری الگوریتم‌ها و ابزارهای وب‌کاوی هستیم [۱].

یکی از قسمت‌های اصلی تشکیل‌دهنده وب‌کاوی، کاوش نحوه استفاده از وب است که خود در بر گیرنده مبحث تحلیل جریان کلیک می‌باشد. هدف اصلی در بخش تحلیل جریان کلیک کشف محتمل‌ترین مسیر حرکت مشتریان در یک سایت است که اهمیت بسیار زیادی در زمینه تجارت الکترونیک و بازاریابی الکترونیکی دارد. برای نمونه می‌توان از آن برای طراحی هرچه بهتر و مؤثرتر وب سایت مورد نظر، به‌منظور حداکثرکردن میزان فروش و سودآوری استفاده کرد.

این مقاله در تاریخ ۶ دی ماه ۱۳۸۵ دریافت و در تاریخ ۱۲ آبان ماه ۱۳۸۸ بازنگری شد.

محمدمهدی سپهری، گروه مهندسی فناوری اطلاعات، دانشگاه تربیت مدرس (email: mehdi.sepehri@modares.ac.ir).

فؤاد مهدوی‌پژوه، کارشناس ارشد مهندسی صنایع (email: mahdavi@okstate.edu).

با توجه به کاربردهای تجاری فراوان پیش‌بینی مسیر حرکت کاربران در یک وب سایت و اهمیت آن در مدیریت ارتباط با مشتری، هدف اصلی در این تحقیق، ارائه مدلی ریاضی بر پایه مدل معروف فروشنده دوره‌گرد جمع‌کننده پاداش برای حل این مسأله می‌باشد.

با این دورنما در بخش ۲ به مرور ادبیات تحلیل جریان کلیک پرداخته‌ایم. بخش ۳ در برگیرنده مدل ریاضی ارائه‌شده می‌باشد و الگوریتم به‌کار رفته برای حل این مدل در بخش ۴ ذکر گردیده است. بخش ۵ شامل نتایج حاصل از پیاده‌سازی مدل بر روی یک وب سایت دانشگاهی است که نتیجه‌گیری و تفسیر نتایج حاصل در بخش ۶ آمده است.

۲- مرور ادبیات تحلیل جریان کلیک

در ادبیات تحلیل جریان کلیک از مدل‌های آماری و احتمالی مختلفی به‌منظور تشخیص الگوهای مسیر حرکت کاربران در وب سایت‌های اینترنتی استفاده شده است.

بستاوروس در سال ۱۹۹۶ و زاکرمن و همکاران در سال ۱۹۹۹ از مدل‌های مارکوفی برای پیش‌بینی صفحات بعدی درخواست شده به‌وسیله کاربران با در نظر گرفتن صفحات قبلی بازدیدشده به‌وسیله آنها استفاده کرده‌اند [۲] و [۳].

هابرمن و همکاران در سال ۱۹۹۸ یک روش قدم‌زنی تصادفی را برای مدل‌سازی تعداد صفحات درخواست شده به‌وسیله کاربران در یک وب سایت خاص به‌کار برده‌اند [۴].

سیدز و همکاران در سال ۲۰۰۰ نشان دادند که مدل‌های سفارش اولیه مارکوف، ابزار توانمندی برای کمک به دسته‌بندی انواع مختلف مشتریان می‌باشند. تمرکز تحقیقات مذکور و بیشتر کارهای انجام‌شده در زمینه علوم کامپیوتر بروی پیش‌بینی با استفاده از الگوریتم‌های پنهان‌کننده، دسته‌بندی کننده یا توصیف‌کننده می‌باشد [۵].

والری در سال ۲۰۰۰ به مدل‌سازی و شبیه‌سازی مشتری پرداخته و بیشتر مدل‌سازی یک مشتری منحصر به فرد مد نظر وی بوده است. این نوع مدل برای پیش‌بینی زمان مشاهده بعدی، کل زمان صرف‌شده برای مشاهده بعدی و زمان صرف‌شده برای مشاهده محصولات متفاوت به‌کار می‌رود. در این مقاله روش‌های مختلف پیش‌بینی نیز مورد بحث و بررسی قرار می‌گیرند [۶].

دیوزینگر و هابر در سال ۲۰۰۰ به توصیف یک مطالعه موردی می‌پردازند که به‌وسیله ASK net و شرکت SAS آلمان انجام شده است که هدف آن تقویت حضور در وب سایت و کسب دانایی در مورد مشتریان می‌باشد [۷].

گلدفارب تقاضای موجود برای ورودی‌های اینترنتی را با استفاده از داده‌های جریان کلیک ۲۶۵۴ کاربر تخمین می‌زند. وی روش گوادگنی و لیتل را برای فهم بهتر انتخاب ورودی‌های اینترنتی به‌کار می‌برد [۸].

سیسمیرو و باکلین در سال ۲۰۰۱ رفتار بازدید مراجعان به یک وب سایت را با استفاده از داده‌های جریان کلیک ذخیره‌شده در فایل‌های ثبت وقایع سرور آن وب، مدل‌سازی کرده‌اند. در این مقاله دو جنبه رفتار بازدید تست شده است:

۱) تصمیم مراجعه‌کنندگان به ادامه بازدید (از طریق ثبت نام کردن یا درخواست صفحات اضافه) و یا تصمیم آنها به خروج از سایت.

۲) مدت زمان صرف‌شده برای بازدید از هر صفحه.

یک مدل دوتایی آماری، تصمیمات هر مراجعه‌کننده را مبنی بر ادامه بازدید یا ترک سایت ارائه می‌کند و یک مدل مخاطرات نسبینیز مدت زمان بازدید هر صفحه را نشان می‌دهد. هر دو مدل می‌توانند به‌صورت جداگانه تخمین زده شوند و یا می‌توانند به یکدیگر پیوند داده شده تا یک بیان مشترک را شکل دهند که به موجب آن می‌توان پذیرفت که تصمیم به ادامه یا خروج از سایت می‌تواند روی مدت زمان بازدید از صفحات اثر بگذارد. مدل به هم پیوسته متشکل از هر دو مدل آماری و تصادفی، برای بررسی تصمیمات بازدید یک نمونه تصادفی ۵۰۰۰ تایی از بازدیدکنندگان سایت یک شرکت اینترنتی فروش خودرو به‌کار رفته است. اطلاعات مورد نیاز در مورد صفحات درخواست‌شده، مدت زمان بازدید صفحات و سایر موارد از فایل‌های ذخیره اطلاعات سرور آن سایت استخراج شده است. مدل مورد نظر با استفاده از روش‌های بیزی تخمین زده شده تا کارآیی لازم را در استفاده از ناهمگونی موضعی موجود در پارامترها داشته باشد [۹].

مو و فیدر در سال ۲۰۰۲ مدلی برای رفتار مشاهده مشتریان اینترنتی بر پایه داده‌های جریان کلیک اینترنتی ارائه می‌کنند. در این مقاله تغییر در رفتار مشاهده مشتریان که با گذشت زمان و با کسب تجربه توسط مشتریان حاصل می‌شود، در نظر گرفته شده است. همچنین رابطه بین تعداد دفعات مشاهده و تمایل به خرید در مشتریان تست شده است و نیز نشان داده شده است که تغییرات موجود در تکرارهای هر مراجعه به‌طور منحصر به فرد اطلاعات بیشتری را با در نظر گرفتن این که کدام بخش از مشتریان احتمال خریدشان بیشتر است، برای ما تأمین می‌کنند [۱۰].

مونتگومری و همکاران در نوامبر ۲۰۰۲ یک مدل احتمالی چندجمله‌ای پویا برای پیش‌بینی مسیری که یک کاربر به هنگام بازدید از یک وب سایت اختیار می‌کند، پیشنهاد کرده‌اند. مدل مورد نظر در یک چهارچوب سلسله مراتبی بیزی برای توجیه ناهمگونی محسوس یا نامحسوس مشتریان فرموله شده است. به علاوه مدل مورد نظر شامل یک پروسه ترکیبی است که حالات چندگانه‌اش به‌وسیله یک مدل تبدیل‌شونده پنهان مارکوف برای تشخیص ناهمگونی کاربران کنترل می‌شود [۱۱].

مونتگومری و فالوتسوس دریافتند که بسیاری از شاخص‌ها و معیارهای رفتار بازدید در طی زمان ثابت هستند [۱۲].

لی و همکاران در سال ۲۰۰۲ یک مدل آماری از رفتار بازدید کاربران به‌وسیله پیش‌بینی تعداد صفحات وب در یک مقوله خاص که به‌وسیله یک کاربر در یک بخش از یک وب مشاهده می‌شوند، ارائه کرده‌اند. هدف از این تحلیل فهم بهتر رفتار مشاهده وب و کمک به پیش‌بینی این است که کدام بخش‌ها منجر به بازدید مجدد و یا خرید می‌شوند. در اینجا از Poisson and discretized tobit models و از تقابل هر دو شکل تک‌متغیره و چندمتغیره این مدل‌ها استفاده شده است. به علاوه چون مجموعه داده‌های مورد نظر دارای ناهمگونی عظیمی در کاربرد بر اساس نوع کاربر و نیز تأکید مناسب بر بازدید می‌باشد، یک مدل جدید چندمتغیره با پروسه مختلطی که حالات چندگانه‌اش با یک صف مارکوف نامحسوس کنترل می‌شوند، پیشنهاد شده است. مشاهده می‌شود که حرکت کاربران بین بخش‌ها که به‌عنوان رفتار بازدید آنها شناخته می‌شود به‌وسیله بررسی تنوع بخش‌هایی که آنها بازدید می‌کنند، قابل تشخیص می‌باشد [۱۳].

جیودیچی و تارانتونا در سپتامبر ۲۰۰۳ مقاله‌ای را با هدف نشان‌دادن این حقیقت که چگونه می‌توان از داده‌های جریان کلیک برای کشف محتمل‌ترین مسیرهای حرکت مشتریان در یک وب سایت استفاده کرد، به چاپ رساندند.

یکی از مدل‌های ارائه‌شده در این مقاله مدل زنجیره‌های مارکوف درجه یک می‌باشد. مدل مذکور دارای دو نقص عمده می‌باشد که عبارتند از:

۱) الگوریتم مورد نظر لزوماً یک مسیر کامل را به‌عنوان جواب به ما نمی‌دهد و ممکن است به حلقه بیافتد.

۲) اگر هم مسیر کاملی به‌دست آید، مسیر به‌دست آمده لزوماً بهینه نیست. یعنی لزوماً برابر با مسیر با بزرگ‌ترین احتمال حاصل از مدل مارکوف نمی‌باشد.

در نهایت کلیه روش‌های ارائه‌شده در این مقاله در دو گروه محلی و سراسری با یکدیگر مقایسه شده‌اند [۱۴].

پارک و فیدر در سال ۲۰۰۴ یک مدل احتمالی برای زمان بازدید بر حسب سایت‌های مختلف، برای درک نحوه به‌کارگیری اطلاعات حاصل از یک سایت برای توجیه رفتار مشتریان هنگام بازدید از سایت‌های دیگر ارائه کرده‌اند. مدل ارائه‌شده به‌طور موفقیت‌آمیزی می‌تواند الگوهای بازدید از یک سایت را رده‌بندی کرده و به ما این امکان را می‌دهد تا اطلاعات بیشتری راجع به تعداد و زمان بازدید مشتریان جدیدی که قبلاً از یک سایت مشخص دیدن نکرده‌اند، به‌دست آوریم. علاوه بر موارد فوق این تحقیق چندین توزیع کاربردی را معرفی می‌کند. به‌ویژه در بخش ادبیات بازاریابی خانواده توزیع‌های چندگانه سارمانوف ارائه شده‌اند. این خانواده با توابع چگالی چندگانه بسیار انعطاف‌پذیر به ما این امکان را می‌دهد تا با حصول اطمینان از این که توزیع‌های کناری به‌طور کامل با چگالی‌های مطلوب تطابق دارند، به یک مدل مختلط دست یابیم. این امر به‌طور فزاینده‌ای تخمین پارامترها، توصیفات مدیریتی و کارایی عملی مدل مورد نظر را افزایش می‌دهد [۱۵].

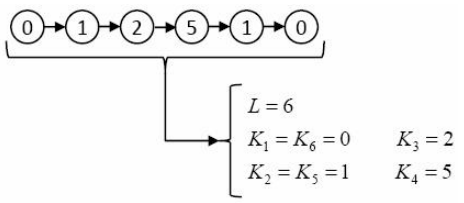
مونتگومری و همکاران در سال ۲۰۰۴ نشان دادند که چگونه اطلاعات مسیر می‌تواند به وسیله یک مدل چندجمله‌ای احتمالی پویا برای بازدید از سایت رده‌بندی و مدل‌سازی شود. مدل ارائه‌شده با استفاده از داده‌های به‌دست آمده از یک فروشنده کتاب عمده اینترنتی تخمین زده شده است. نتایج حاصل نشان می‌دهند که اجزای تشکیل‌دهنده حافظه مدل در پیش‌بینی دقیق یک مسیر بسیار تأثیرگذار هستند. یکی از کاربردهای بالقوه مدل ارائه‌شده، پیش‌بینی انجام خرید می‌باشد [۱۶].

۳- مدل‌سازی

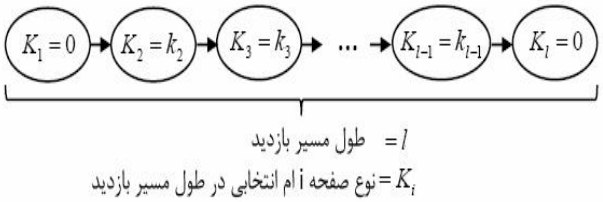
۳-۱ انتخاب مدل مناسب برای حل مسأله تحقیق

بعد از مطالعه ادبیات موضوع تحلیل جریان کلیک، مدل زنجیره‌های مارکوف به‌عنوان مدلی مناسب برای حل مسئله تحقیق انتخاب گردید. در مقالاتی که در آنها از مدل‌های مارکوفی برای پیش‌بینی مسیر حرکت یک کاربر استفاده شده است، محققان اشاره کرده‌اند که صفحه kام انتخاب‌شده به‌وسیله فرد در مسیر بازدید وی از یک وب سایت، اساساً وابسته به محتوا و خصوصیات موجود در صفحه $k-1$ام انتخاب‌شده به‌وسیله وی می‌باشد که این خود نشان‌دهنده صحت استفاده از مدل‌های مارکوف درجه یک است [۱۲]، [۱۳] و [۱۶].

در ادامه با در نظر گرفتن مدل مارکوف درجه یک به‌عنوان مدلی



شکل ۱: پیشامد بازدید کاربری که به‌ترتیب صفحات شماره ۱، ۲، ۵ و ۱ را مشاهده کرده است.



شکل ۲: پیشامد بازدید برای یک کاربر خاص در حالت کلی.

مناسب برای مدل‌سازی این مسأله، به تعریف پیشامد بازدید یک کاربر از یک سایت با استفاده از این مدل می‌پردازیم. با در نظر گرفتن وب سایتی که دارای N صفحه متمایز است، این پیشامد شامل توالی صفحات مشاهده‌شده به‌وسیله کاربر در یک بازدید است که ممکن است در این مسیر بازدید، یک صفحه مشخص چندین بار مشاهده شود.

برای نشان‌دادن پیشامد رفتار بازدید در این مقاله از متغیرهای تصادفی $K_1,K_2,\ldots,K_L$ استفاده می‌کنیم که اندیس L در آن نشان‌دهنده تعداد صفحات مشاهده‌شده به‌وسیله وی می‌باشد.

با فرض مشاهده L صفحه از وب سایت توسط کاربر مورد نظر، متغیرهای تصادفی $K_1,K_2,\ldots,K_L$ نشان‌دهنده نوع صفحات انتخابی در مسیر طی‌شده به‌وسیله کاربر در یک بار بازدید وی می‌باشند.

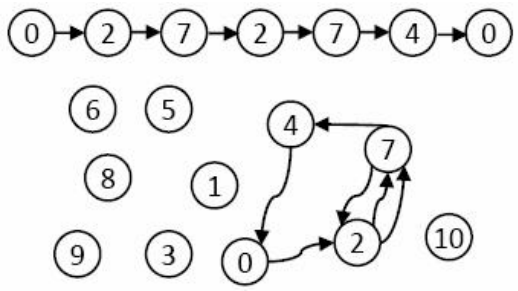
در این مقاله برای نشان‌دادن هر پیشامد مسیر حرکت یک کاربر، علاوه بر نمایش توالی صفحات طی‌شده به‌وسیله وی، برای نمایش ورود و خروج به وب سایت از یک گره مجازی (۰) استفاده می‌شود که گره بعد از آن در ابتدای مسیر، نشان‌دهنده اولین صفحه مشاهده‌شده به‌وسیله کاربر و گره قبل از آن در انتهای مسیر، نشان‌دهنده آخرین صفحه مشاهده‌شده به‌وسیله او می‌باشند. به این ترتیب برای هر پیشامد مسیر بازدید یک کاربر از سایت داریم

$$K_1 = K_L = 0 \qquad (1)$$

که در آن L برابر تعداد صفحات مشاهده‌شده به‌وسیله وی از سایت در این بازدید به اضافه ۲ (دو گره مجازی صفر) می‌باشد. اگر کاربر وارد وب سایت مذکور شده و به‌ترتیب صفحات ۱، ۲، ۵ و ۱ را مشاهده کرده و از وب سایت خارج شود، در این مقاله پیشامد بازدید وی مطابق با شکل ۱ نمایش داده می‌شود.

با توجه به تعاریف ارائه‌شده در قسمت قبل، بدیهی است که همواره به‌ازای هر بازدید از سایت داریم: $L \geq 3$.

در برخی از وب سایت‌ها که تعداد صفحات تشکیل‌دهنده آنها بسیار زیاد است، می‌توان به‌جای در نظر گرفتن تک‌تک صفحات سایت در مدل، کلیه صفحات را به چند بخش اصلی تشکیل‌دهنده آنها افراز کرده و از این بخش‌ها به‌عنوان جایگزین صفحات در مدل استفاده کرد [۱۲]، [۱۳]، [۱۶] و [۱۷].



شکل ۳: مسیر و گراف حرکت کاربری که به‌ترتیب صفحات ۲، ۷، ۲، ۷ و ۴ را بازدید کرده است.

در این حالت اندیس L نشان‌دهنده تعداد بخش‌های مشاهده‌شده به‌وسیله کاربر در یک بار بازدید وی از وب سایت و پارامتر N نشان‌دهنده تعداد کل بخش‌های تشکیل‌دهنده وب سایت مذکور می‌باشند.

با در نظر گرفتن تعاریف ذکرشده برای رفتار بازدید، برای مدل‌سازی تابع هدف، پیشامد کلی شکل ۲ را برای یک کاربر خاص در نظر گرفته و رابطه احتمالی آن را می‌نویسیم

$$\begin{aligned} &P(A) = \\ &P(\{K_1 = 0, K_2 = k_2, K_3 = k_3, \ldots, K_{l-1} = k_{l-1}, K_l = 0\}) \\ &= p(0 \mid k_{l-1}) \times p(k_{l-1} \mid k_{l-2}) \times \ldots \times p(k_3 \mid k_2) \times p(k_2 \mid 0) \end{aligned} \qquad (2)$$

با توجه به رابطه احتمالی به‌دست آمده، هدف ما در این مقاله برای حل مسئله تحقیق، یافتن مقدار متغیرهای تصادفی K_1 تا K_l برای یک کاربر مشخص با استفاده از روش برنامه‌ریزی صفر و یک به قسمی است که مقدار $P(A)$ حداکثر شود. بنابراین (۳) به‌دست می‌آید.

۳-۲ مدل ریاضی پیش‌بینی مسیر حرکت یک کاربر

در ادامه برای یافتن $\max_{\{K_1,K_2,\ldots,K_L\}} P(K)$ با استفاده از برنامه‌ریزی صفر و یک، ابتدا تابع هدف $P(K)$ را که با استفاده از مدل زنجیره مارکوف درجه یک مدل‌سازی می‌شود، با به‌کارگیری تبدیل مناسب به تابع هدف روش برنامه‌ریزی صفر و یک تبدیل می‌کنیم. سپس محدودیت‌های این مدل برنامه‌ریزی صفر و یک را نوشته و مدل نهایی را می‌یابیم.

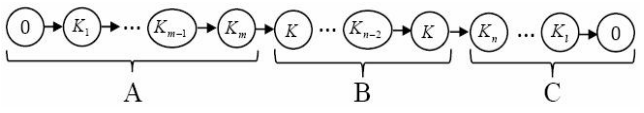
۳-۲-۱ مدل‌سازی تابع هدف برنامه‌ریزی صفر و یک

برای نمونه وب سایتی را در نظر بگیرید که دارای ۱۰ صفحه (یا بخش) متمایز باشد. یک پیشامد مسیر حرکت موجه در شکل زیر نمایش داده شده است. در این پیشامد فردی در یک بار بازدید خود از این وب سایت، ابتدا وارد سایت شده و صفحه شماره ۲ را به‌عنوان اولین صفحه از این وب سایت مشاهده می‌کند. سپس به‌ترتیب صفحات ۷، ۲، ۷ و ۴ را دیده و از سایت خارج می‌شود. مسیر و گراف حرکت این کاربر معادل شکل ۳ می‌باشد.

برای مدل‌سازی تابع هدف این مسئله، با فرض این که پارامتر $N = n$ نشان‌دهنده تعداد صفحات (بخش‌های) وب سایت مورد نظر باشد، متغیر عدد صحیح x_{ij} را برابر با تعداد دفعات حرکت کاربر ار صفحه (بخش) i به صفحه (بخش) j تعریف می‌کنیم. بنابراین در مثال شکل ۳ داریم

$$\begin{aligned} &x_{02} = 1 \;,\; x_{27} = 2 \;,\; x_{72} = 1 \;,\; x_{74} = 1 \;,\; x_{40} = 1 \\ &\text{for all other } i,j \quad x_{ij} = 0 \end{aligned} \qquad (4)$$

$$\begin{aligned} &\max\ P(A) = \max_{\{K_1,K_2,\ldots,K_l\}} P(K) = \max_{\{K_1,K_2,\ldots,K_l\}} P(\{K_1 = 0, K_2 = k_2, \ldots, K_{l-1} = k_{l-1}, K_l = 0\}) \\ &= \max_{\{K_1,K_2,\ldots,K_l\}} p(0 \mid k_{l-1}) p(k_{l-1} \mid k_{l-2}) \ldots p(k_3 \mid k_2) p(k_2 \mid 0) \end{aligned} \qquad (3)$$



شکل ۴: پیشامد بازدید عمومی (S) که دارای حداقل یک گره تکراری k است.

به این ترتیب ابتدا به نظر می‌رسد که مسئله ما یک مسأله برنامه‌ریزی عدد صحیح می‌باشد. اما یک ویژگی بسیار مهم در مدل‌های مارکوفی درجه یک، مدل ما را به یک مدل برنامه‌ریزی صفر و یک تبدیل می‌کند. ویژگی مذکور به شرح زیر است:

در مدل مارکوف درجه یک، هر مسیر دارای گره تکراری (به‌جز گره صفر) قابل تبدیل به مسیر بدون گره تکراری (به‌جز گره صفر) با مقدار تابع هدف (احتمال وقوع) بیشتر می‌باشد.

برای اثبات این ادعا شکل ۴ را در نظر بگیرید که پیشامد بازدید عمومی (S) را که حداقل دارای یک گره تکراری k است، نشان می‌دهد. این پیشامد بازدید عمومی را می‌توان به سه بخش اصلی افراز کرد. بخش A که در بر گیرنده توالی صفحات طی‌شده در این پیشامد از ابتدا تا قبل از اولین تکرار گره k است. بخش B که در برگیرنده خود دو تکرار گره k و تمامی گره‌های بین این دو تکرار است و در نهایت بخش c که در برگیرنده توالی صفحات مشاهده‌شده بعد از دومین تکرار گره k تا انتهای این پیشامد بازدید عمومی است.

احتمال رخدادن پیشامد عمومی S به شرح زیر است

$$p_S = \underbrace{p(\cdot|k_l)\dots p(k_n|k)}_{p_C} \times \underbrace{p(k|k_{n-1})\dots p(k|k_m)}_{p_B} \times \underbrace{p(k_m|k_{m-1})\dots p(k_1|\cdot)}_{p_A} \quad (۵)$$

حال اگر قسمت B را که در بر گیرنده هر دو گره تکراری و کلیه گره‌های بین آنهاست از پیشامد عمومی S حذف کرده و فقط یک گره k را جایگزین آن کنیم پیشامد S' حاصل می‌شود که در شکل ۵ نمایش داده شده است.

احتمال رخدادن پیشامد S' برابر است با

$$p_{S'} = \underbrace{p(\cdot|k_l)\dots p(k_n|k)}_{p_C} \times p(k|k_m) \times \underbrace{p(k_m|k_{m-1})\dots p(k_1|\cdot)}_{p_A} \quad (۶)$$

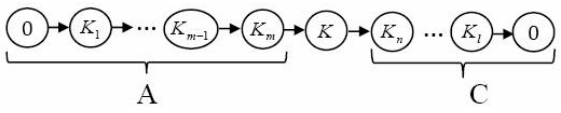
با توجه به این که احتمال همواره مقداری بین صفر و یک دارد رابطه زیر همواره برقرار است

$$p(k|k_m) > \underbrace{p(k|k_{n-1})\dots p(k|k_m)}_{p_B} \quad (۷)$$

پس با توجه به (۵) تا (۷) همواره داریم $p_{S'} > p_S$.

همان‌طور که دیدیم با حذف دو گره تکراری و تمامی گره‌های بین آنها و جایگزینی یک گره از همان نوع به جای آنها پیشامدی با احتمال وقوع بیشتر حاصل شد. حال اگر به‌ازای تمام گره‌های تکراری این پیشامد، این کار را انجام دهیم در نهایت یک پیشامد بازدید بدون گره تکراری با مقدار احتمال وقوع بیشتری نسبت به تمام پیشامدهای قبلی حاصل می‌شود و ویژگی مذکور به اثبات رسیده است.

با توجه به این نکته در می‌یابیم که جواب‌ها با مقدار x_{ij} بزرگ‌تر از یک، گرچه ممکن است موجه باشند اما هرگز بهینه نیستند. بنابراین برای یافتن جواب بهینه مسئله برنامه‌ریزی عدد صحیح قبل کافی است جواب بهینه مسئله برنامه‌ریزی صفر و یک جدید را یافت چرا که جواب بهینه



شکل ۵: پیشامد بازدید (S').

مسئله برنامه‌ریزی صفر و یک جدید قطعاً جواب بهینه مسئله برنامه‌ریزی عدد صحیح قبل هم می‌باشد.

در این حالت، متغیر تصمیم جدید x_{ij} در صورتی که کاربر از صفحه (بخش) i به صفحه (بخش) j رفته باشد، برابر یک و در غیر این صورت برابر صفر خواهد بود.

برای محاسبه تابع هدف $P(K)$ برای این پیشامد جدید با استفاده از مدل زنجیره مارکوف درجه یک داریم

$$P(K) = P(K_1=0, K_2=2, K_3=7, K_4=4, K_5=0) = p'_k \quad (۸)$$

با استفاده از مدل زنجیره مارکوف درجه یک، احتمال پیشامد بالا برابر است با

$$\begin{aligned} p'_k &= p(\cdot|4)p(4|7)p(7|2)p(2|\cdot) \Rightarrow \\ \ln p'_k &= \ln(p(\cdot|4)) + \ln(p(4|7)) \\ &\quad + \ln(p(7|2)) + \ln(p(2|\cdot)) \Rightarrow \\ \ln p'_k &= x_{40}\ln(f(\cdot|4)) + x_{74}\ln(f(4|7)) \\ &\quad + x_{27}\ln(f(7|2)) + x_{02}\ln(f(2|\cdot)) \end{aligned} \quad (۹)$$

بنابراین در حالت کلی برای هر توالی ممکن (موجه) از صفحات بازدیدشده می‌توان تابع هدف صفر و یک زیر را تعریف نمود

$$\begin{aligned} F &= \ln P(K) = \ln p'_k \\ &= \sum_{j=0}^{n}\sum_{i=0}^{n} x_{ij}\ln(f(j|i)) \xrightarrow{\ln(f(j|i))=p_{ij}} \\ F &= \ln p'_k = \sum_{j=0}^{n}\sum_{i=0}^{n} p_{ij}x_{ij} \Rightarrow \\ \max F &\equiv \min -F \Rightarrow \\ \min -F &= \min -\sum_{j=0}^{n}\sum_{i=0}^{n} p_{ij}x_{ij} \xrightarrow{c_{ij}=-p_{ij}} \\ \min -F &= \min \sum_{j=0}^{n}\sum_{i=0}^{n} c_{ij}x_{ij} \end{aligned} \quad (۱۰)$$

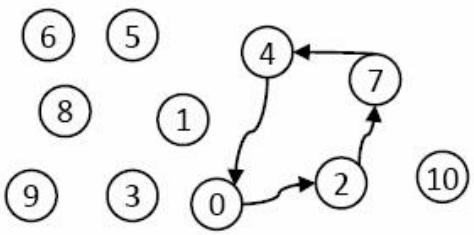
بنابراین با فرض $c_{ij} = -\ln(f(j|i))$ تابع هدف برنامه‌ریزی صفر و یک ما برابر با $\min F = \sum_{j=0}^{n}\sum_{i=0}^{n} c_{ij}x_{ij}$ است. لازم به ذکر است که منظور از $f(j|i)$، درایه واقع در سطر iام و ستون jام ماتریس انتقال[1] مدل مارکوف درجه یک برای کاربر مورد نظر است که برابر با احتمال رفتن از صفحه iام به صفحه jام می‌باشد و با استفاده از فراوانی نسبی حرکت کاربر از صفحه iام به صفحه jام نسبت به کل حرکات او در لاگ فایل‌های سرور وب سایت مورد نظر محاسبه می‌شود.

۳-۲-۲ مدل‌سازی محدودیت‌های برنامه‌ریزی صفر و یک

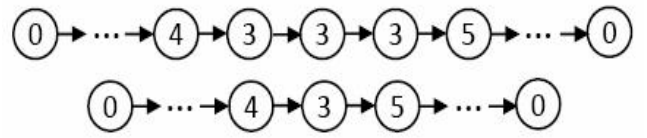
با در نظر گرفتن گراف مسیر طی‌شده در مثال قبل در شکل ۶ می‌توان کلیه محدودیت‌های این مسئله برنامه‌ریزی صفر و یک را به شکل زیر بیان نمود.

۱) در هیچ یک از مسیرهای بازدید موجه این مسئله، یک صفحه خاص چند بار پشت سر هم بلافاصله مشاهده نمی‌شود. در واقع تکرار چند بار پشت سر هم و بلافاصله یک صفحه خاص در یک

1. Transition Matrix



شکل ۶: گراف مسیر طی‌شده توسط کاربر قبل بعد از تبدیل مسئله عدد صحیح به مسئله صفر و یک.



شکل ۷: در هیچ یک از مسیرهای موجه بازدید در این مسئله حلقه (لوپ) به طول صفر نداریم.

مشاهده بازدید را یک بار مشاهده آن در نظر می‌گیریم. در واقع در گراف مسیر بازدید، لوپ (حلقه) به طول صفر نداریم (شکل ۷). پس در واقع داریم

$$x_{ij} = ۰ \quad \text{if} \quad i = j \qquad (۱۱)$$

۲) در طول مسیر بازدید به هر گره که وارد می‌شویم باید بتوانیم از آن خارج شویم

$$\sum_{i=۰}^{n} x_{ik} = \sum_{j=۰}^{n} x_{kj} \quad , \quad k = ۰,۱,\ldots,n \qquad (۱۲)$$

۳) هر پیشامد مسیر بازدید از گره مجازی صفر شروع و به آن هم ختم می‌شود

$$\sum_{i=۰}^{n} x_{i۰} = ۱ \qquad (۱۳)$$

محدودیت زیر که همان محدودیت شروع از گره مجازی صفر است، خودبه‌خود با در نظر گرفتن هر دو محدودیت شماره ۲ و ۳ با هم، همواره برقرار می‌باشد

$$\sum_{j=۰}^{n} x_{۰j} = ۱ \qquad (۱۴)$$

۴) در هیچ یک از مسیرهای موجه برای این مسئله برنامه‌ریزی صفر و یک زیر تور بدون گره صفر نداریم. شکل ۸ مثالی از یک جواب غیر موجه که دارای دو زیر تور بدون گره صفر است را نشان می‌دهد. محدودیت‌های حذف زیر تور به‌صورت زیر بیان می‌شوند [۱۸] تا [۲۰]

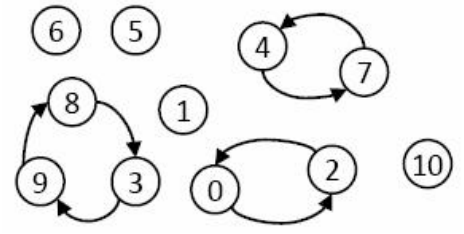
$$\begin{aligned} & y_i = \sum_{j=۰}^{n} x_{ij} && \text{for all } i \\ & \sum_{i\in k'} \sum_{j\in k} x_{ij} \geq y_h && S = \{۰,۱,۲,\ldots,n\} \\ & && \text{for each } h \in K', \\ & && K \subset S \text{ and } ۰ \in K, \\ & && K' = S - K \end{aligned} \qquad (۱۵)$$

۵) محدودیت حذف کلیه جواب‌ها با گره‌های تکراری به شرح زیر است

$$y_i = \{۰,۱\} \quad \text{for all } i \qquad (۱۶)$$

۶) در نهایت هم کلیه متغیرهای x_{ij} از نوع عدد صفر و یک می‌باشند

$$x_{ij} = \{۰,۱\} \quad \text{for all } i,j \qquad (۱۷)$$



شکل ۸: جواب غیر موجهی که دارای دو زیر تور بدون گره صفر می‌باشد.

۳-۲-۳ مدل برنامه‌ریزی صفر و یک نهایی

مدل برنامه‌ریزی صفر و یک نهایی به‌صورت زیر می‌باشد

$$\min \sum_{i=۰}^{n} \sum_{j=۰}^{n} c_{ij} x_{ij}$$

s.t.

$$\begin{aligned} & x_{ij} = ۰ && \text{if } i = j \\ & \sum_{i=۰}^{n} x_{ik} = \sum_{j=۰}^{n} x_{kj} && k = ۰,۱,\ldots,n \\ & \sum_{i=۰}^{n} x_{i۰} = ۱ && \\ & y_i = \sum_{j=۰}^{n} x_{ij} && \text{for all } i \\ & \sum_{i\in k'} \sum_{j\in k} x_{ij} \geq y_h && \begin{cases} \text{for each } h \in K' \\ S = \{۰,۱,۲,\ldots,n\} \\ K \subset S \text{ and } ۰ \in K \\ K' = S - K \end{cases} \\ & x_{ij} = \{۰,۱\} && \text{for all } i,j \\ & y_i = \{۰,۱\} && \text{for all } i \end{aligned}$$

مدل ارائه‌شده حالت خاصی از مسئله فروشنده دوره‌گرد جمع‌کننده پاداش (PCTSP) می‌باشد که در ادامه به تشریح آن مدل می‌پردازیم.

۳-۳ مسئله فروشنده دوره‌گرد جمع‌کننده پاداش با فرض شروع و خاتمه مسیر در گره صفر[1] (PCTSP)

مسئله فروشنده دوره‌گرد جمع‌کننده پاداش (PCTSP) با فرض شروع و خاتمه مسیر در گره صفر، مسئله فروشنده دوره‌گردی است که در آن فروشنده مذکور از شهر صفر شروع کرده و در انتهای مسیر خود نیز مجدداً به این شهر باز می‌گردد ولی بر خلاف مسئله TSP لازم نیست در طول مسیر حرکت خود از تمامی شهرها عبور کند و همچنین ممکن است از بعضی از شهرها بیش از یک بار عبور کند. عبور از هر شهر برای فروشنده مذکور دارای پاداشی است که مقدار آن از قبل مشخص است.

در این مسأله هدف اصلی تعیین توری است که در حالی که پاداش جمع‌آوری شده از یک حد پایینی بالاتر است، هزینه کل سفر را مینیمم کند [۱۸] تا [۲۰].

مدل ریاضی مسئله PCTSP با فرض شروع و خاتمه مسیر در گره صفر، در حالی که هدف مینیمم‌کردن هزینه سفر به‌شرطی است که پاداش کل جمع‌آوری شده از یک حد پایینی بزرگ‌تر شود به شکل زیر می‌باشد:

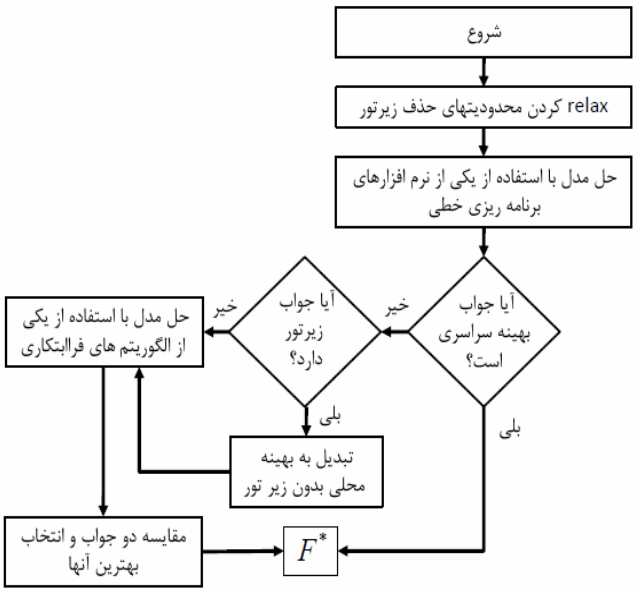
c_{ij}: هزینه سفر از شهر i به شهر j.

p_i: پاداش حاصل از عبور از شهر i.

g: حد پایین پاداش جمع‌آوری شده.

x_{ij}: متغیر تصمیم صفر و یک است که اگر فروشنده در طی مسیر خود از شهر i به شهر j سفر کند مقدارش یک و در غیر این

1. Prize Collecting Traveling Salesman Problem



شکل ۹: الگوریتم پیشنهادی برای حل مدل.

صورت صفر می‌باشد.

y_i: متغیر صفر و یکی است که اگر شهر i در طی مسیر فروشنده توسط وی بازدید شود مقدارش یک و در غیر این صورت صفر می‌باشد

$$\min \sum_{i=0}^{n}\sum_{j=0}^{n} c_{ij}x_{ij}$$

s.t.

$$x_{ij} = 0 \qquad \text{if } i = j$$

$$\sum_{i=0}^{n} x_{ik} = \sum_{j=0}^{n} x_{kj} \qquad k = 0,1,\ldots,n$$

$$\sum_{i=0}^{n} x_{i0} = 1$$

$$\sum_{i=0}^{n} p_i \sum_{j=0}^{n} x_{ij} \geq g$$

$$y_i = \sum_{j=0}^{n} x_{ij} \qquad \text{for all } i$$

$$\sum_{i\in k'}\sum_{j\notin k} x_{ij} \geq y_h \qquad \begin{cases} \text{for each } h \in K' \\ S = \{0,1,2,\ldots,n\} \\ K \subset S \text{ and } 0 \in K \\ K' = S - K \end{cases}$$

$$x_{ij} = \{0,1\} \qquad \text{for all } i,j$$

$$y_i = \{0,1\} \qquad \text{for all } i$$

مسئله فروشنده دوره‌گرد جمع‌کننده پاداش (TSSP)، یک مسئله Np - hard بوده و مشاهده می‌شود که تعداد محدودیت‌های حذف زیر تور آن $\left[L!(n+1-L)!\right]/\sum_{L=2}^{n+1}(n+1-L)(n+1)!$ با اضافه‌شدن تعداد گره‌ها (صفحات) به‌طور انفجارآمیزی افزایش می‌یابد [۱۸] تا [۲۰].

حالت خاصی از مسئله فروشنده دوره‌گرد جمع‌کننده پاداش با فرض شروع و خاتمه مسیر در گره صفر را در نظر بگیرید که پارامترهای آن به شکل $p_i = 1$ و $g = 2$ تعریف شده‌اند.

با واردکردن مقادیر این پارامترها در مدل این مسأله، مدلی دقیقاً به شکل مدل برنامه‌ریزی صفر و یک نهایی پیشنهادی ما برای پیش‌بینی مسیر حرکت یک کاربر در یک وب سایت در قسمت قبل حاصل می شود. لازم به ذکر است که محدودیت $\sum_{i=0}^{n}\sum_{j=0}^{n} x_{ij} \geq 2$، با توجه به وجود دو محدودیت $\sum_{i=0}^{n} x_{ik} = \sum_{j=0}^{n} x_{kj}$ و $\sum_{j=0}^{n} x_{ij} = 1$ محدودیت زایدی بوده و از مسئله حذف می‌شود.

بنابراین مدل پیشنهادی ما حالت خاصی از مسئله NP Hard فروشنده دوره‌گرد جمع‌کننده پاداش با فرض شروع و خاتمه مسیر در گره صفر است. پس خود مدل ارائه‌شده هم یک مسئله NP Hard می‌باشد.

٤- الگوریتم پیشنهادی برای حل مدل

برای حل مدل پیشنهادی ابتدا محدودیت‌های حذف زیر تور آن را حذف[1] کرده سپس مسئله جدید حاصل را با استفاده از نرم‌افزارهای برنامه‌ریزی خطی حل می‌کنیم.

نکته بسیار مهمی که باید در حل این مسئله جدید (relax شده) در نظر گرفت این است که جواب بهینه قطعی آن، هرگز زیر تور ندارد. برای اثبات این ادعا از برهان خلف استفاده می‌کنیم. فرض کنید جواب بهینه قطعی برای این مسأله (relax شده) به‌دست آمده که حداقل دارای یک زیر تور است. حال جوابی را در نظر بگیرید که از حذف زیر تور مذکور از جواب بهینه قطعی قبلی حاصل می‌شود. این جواب، پاسخ موجه بهتری نسبت به بهینه قطعی قبلی است زیرا اولاً با توجه به محدودیت‌های موجود موجه است ثانیاً هزینه‌اش کمتر است. پس جواب بهتری نسبت به بهینه قطعی مسئله به‌دست می‌آید که خود یک تناقض است.

توجه به این نکته هم ضروری است که با حل مسئله جدید (relax شده) اگر بهینه محلی با زیر تور حاصل شود، با حذف کلیه زیر تورهای موجود در آن به‌راحتی قابل تبدیل به بهینه محلی بدون زیر تور با تابع هدف بهتر می‌باشد زیرا جواب جدید علاوه بر موجه‌بودن هزینه کمتری هم دارد.

در صورت به‌دست آمدن بهینه محلی در حل با نرم‌افزار برنامه‌ریزی خطی، مسئله مورد نظر را با استفاده از یکی از الگوریتم‌های فرابتکاری حل و بعد از مقایسه دو جواب، بهترین را انتخاب می‌نماییم.

با توجه به نکته ذکرشده، الگوریتم ارائه‌شده در شکل ۹ را برای حل مدل پیش‌بینی مسیر حرکت یک کاربر در یک سایت ارائه می‌کنیم.

لازم به ذکر است که در الگوریتم ارائه‌شده از نرم‌افزار LINGO برای حل مدل برنامه‌ریزی صفر و یک استفاده شده است. الگوریتم فرابتکاری شبیه‌سازی بازپختی[2] هم با توجه به کاربرد فراوان آن در ادبیات برای حل مسئله فروشنده دوره‌گرد، به‌عنوان الگوریتم فرابتکاری مناسب برای این حل مدل برنامه‌ریزی صفر و یک ذکر شده در نظر گرفته شده است [۲۱] تا [۲۳].

٥- پیاده‌سازی مدل ارائه‌شده برای وب سایت دانشگاه تربیت مدرس و حل آن

آدرس وب سایت دانشگاه تربیت مدرس، www.modares.ac.ir می‌باشد. از آنجا که این وب سایت دارای تعداد صفحات بسیار زیاد با زیر شاخه‌های فراوان است، می‌توان به‌جای در نظر گرفتن تک‌تک صفحات سایت در مدل، کلیه صفحات را به چند بخش اصلی تشکیل‌دهنده آنها افراز کرده و از این بخش‌ها به‌عنوان جایگزین صفحات در مدل استفاده کرد. کلیه بخش‌های تشکیل‌دهنده وب سایت دانشگاه تربیت مدرس به ۱۰ بخش اصلی افراز گردید که این بخش‌ها و کد اختصاص‌یافته به هر یک به ترتیب عبارتند از:

۱) Home Page

۲) News & Events

1. Relax
2. Simulated Annealing

```
#Software: Microsoft Internet Information Services 5.0
#Version: 1.0
#Date: 2005-01-01 00:00:21
#Fields: date time c-ip cs-username s-sitename s-computername s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status sc-bytes cs-bytes time-taken cs-version cs-host cs(User-Agent) cs(Referer)
2005-01-01 00:00:21 68.202.72.28 - W3SVC3 WEBSERVER 213.176.28.7 80 GET /jast/GB/default.asp PagePosition=12 200 0 457 78 HTTP/1.1 www.modares.ac.ir Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+Q312461;+SV1;+.NET+CLR+1.0.3705)
http://www.google.com/search?q=youth+nudist+photos&hl=en&lr=&start=20&sa=N
2005-01-01 00:00:23 68.202.72.28 - W3SVC3 WEBSERVER 213.176.28.7 80 GET /jast/GB/default_style_css.asp - 200 0 388 62 HTTP/1.1 www.modares.ac.ir Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+Q312461;+SV1;+.NET+CLR+1.0.3705) http://www.modares.ac.ir/jast/GB/default.asp?PagePosition=12
```

شکل ۱۰: بخشی از یکی از لاگ فایل‌های سرور وب سایت دانشگاه تربیت مدرس.

۳) General Information
۴) Research
۵) Faculties & People
۶) Facilities & Services
۷) Site Map
۸) Search
۹) Contact Us
۱۰) Link to Other Websites

کلیه لاگ فایل‌های سرور این وب سایت به شکل فایل‌های متنی[1] در اتاق سرور سایت مرکزی دانشگاه موجود می‌باشد که برای این مدل‌سازی لاگ فایل‌های مربوط به استفاده یک سال کاربران از این مرکز دریافت گردید. بخشی از یکی از لاگ فایل‌های سرور این سایت که مربوط به روز اول ژانویه ۲۰۰۵ است، در شکل ۱۰ آمده است.

هر بار کاربری وارد این وب سایت شده و روی هر لینک از این سایت کلیک کند، سرور این وب سایت یک خط در لاگ فایل مورد نظر ثبت می‌کند. برخی از قسمت‌های[2] مهم اطلاعات موجود در هر خط ثبت شده در این لاگ فایل‌ها عبارتند از:

۱) تاریخ[3]: تاریخ روزی است که کاربر روی لینک مورد نظر کلیک کرده است.
۲) زمان[4]: ساعت، دقیقه و ثانیه‌ای است که کاربر روی لینک مورد نظر کلیک کرده است.
۳) شماره ip کاربر (c-ip (client ip)): شماره ip رایانه‌ای است که کاربر از طریق آن روی لینک مورد نظر کلیک کرده است.
۴) زمینه cs-uri-stem: نشان‌دهنده نام صفحه‌ای که این لینک کلیک شده توسط کاربر مستقیماً وی را به آن صفحه منتقل می‌کند و در واقع اگر عبارت ثبت‌شده در این قسمت به دنباله عبارت www.modares.ac.ir اضافه شود نشان‌دهنده آدرس اینترنتی صفحه درخواست شده به‌وسیله کاربر می‌باشد.

یکی از محدودیت‌های موجود در این تحقیق، عدم وجود نام کاربری[5] برای کاربران این وب سایت است چرا که این سایت طوری طراحی نشده است که کاربران قبل از ورود به آن مجبور باشند از طریق نام کاربری و کلمه رمز[6] مخصوص خود به آن وارد[7] شوند. در این تحقیق با توجه به این که فرض اختصاص یک به یک هر شماره ip به هر کاربر، تأثیری در کلیت مدل و اجرای آن ندارد، با قبول این محدودیت فرض می‌کنیم که هر شماره ip دقیقاً نشان‌دهنده یک کاربر خاص می‌باشد.

در ادامه ۲۰ شماره ip مختلف را به‌عنوان ۲۰ کاربر مختلف در نظر گرفته و بعد از محاسبه ماتریس انتقال ۱۱×۱۱ هر یک با استفاده از اطلاعات موجود در لاگفایل‌ها و روش فراوانی نسبی، محتمل‌ترین مسیر حرکت هر یک را پیش‌بینی نمودیم. برای نشان‌دادن قدرت پیش‌بینی مدل پیشنهادی خود، یک بار دیگر مسیر حرکت هر یک از این ۲۰ کاربر مذکور را با استفاده از الگوریتم پیشنهادی توسط P.Giudici در مقاله Web Mining Pattern Discovery پیش‌بینی کردیم و احتمال رخ‌دادن جواب‌های حاصل را با هم مقایسه نمودیم. خلاصه جواب‌های حاصل مطابق جدول ۱ می‌باشد.

منظور از ستون نوع در جدول یادشده، سراسری یا محلی‌بودن جواب حاصل از الگوریتم پیشنهادی است که حرف G به معنی بهینه سراسری و حرف L به معنی بهینه محلی است.

همان‌طور که مشاهده می‌شود، مدل پیشنهادی در ۱۰۰ درصد موارد جواب بهینه سراسری به‌دست می‌دهد. همچنین با به‌کارگیری مدل پیشنهادی، در ۹۰ درصد موارد جواب‌های بهتری نسبت به الگوریتم P.Giudici به‌دست می‌آید و در ۱۰ درصد باقیمانده هم جواب حاصل از دو روش یکسان هستند. لازم به ذکر است که الگوریتم P.Giudici در ۶۵ درصد موارد به لوپ افتاده و مسیر کاملی به‌دست نمی‌دهد.

با توجه به نتایج حاصل می‌توان دریافت که مدل پیشنهادی روشی بسیار بهتر و با جواب‌های دقیق‌تری نسبت به روش P.Giudici می‌باشد.

۶- نتیجه‌گیری

بسیار واضح و روشن است که رقابت در بازارهای تجاری چند سال آینده برای کسب سود بیشتر در فضای مجازی وب‌ها بوده و تاجرانی موفق به جذب مشتریان بیشتر و بقا در بازارهای تجاری اینترنتی خواهند بود که هرچه بیشتر و بهتر مشتریان خود را شناخته، با آنها ارتباط برقرار کرده و نیاز آنها را به بهترین شکل ممکن پاسخ دهند و به‌طور خلاصه ارتباط با مشتری را به بهترین وجه ممکن مدیریت کنند.

برای مدیریت موفق ارتباط با مشتری، مدیران باید قادر باشند تا داده‌ها و اطلاعات موجود در لاگ فایل‌های وب سرور خود را که حاصل از بازدیدهای مشتریان از وب آنهاست، به‌خوبی با استفاده از ابزارهای وب‌کاوی که تحلیل جریان کلیک یکی از آنهاست، تحلیل کرده و دانش‌های بسیار ارزشمند نهفته در این داده‌ها را استخراج کرده و از آن برای بازاریابی و برقراری رابطه هرچه موفق‌تر با مشتریان استفاده کنند. با این هدف، در این مقاله برآنیم تا با ارائه مدلی ریاضی بر پایه مدل فروشنده دوره‌گرد گردآورنده جایزه، ابزاری برای پیش‌بینی مسیر حرکت یک کاربر خاص در یک وب سایت ارائه کرده تا بتواند تاجران اینترنتی را هرچه بیشتر و مؤثرتر در مدیریت ارتباط با مشتری (CRM) یاری دهد.

برای نمایش کارآیی مدل مذکور، مسیر حرکت ۲۰ کاربر مختلف یک وب سایت دانشگاهی با استفاده از این مدل پیش‌بینی گردید و نشان داده شد که مدل پیشنهادی در ۹۰ درصد موارد دقیق‌تر از مدل رایج P.Giudici عمل می‌کند و در ۱۰ درصد باقیمانده، عملکرد این دو مدل یکسان می‌باشد. یکی دیگر از نقاط قوت مدل پیشنهادی نسبت به مدل‌های رایج مارکوفی این است که برخلاف این مدل‌ها، مدل برنامه‌ریزی صفر و یک ارائه‌شده در این تحقیق هرگز به حلقه نمی‌افتد و همواره جواب کامل به‌دست می‌دهد.

1. Text
2. Fields
3. Date
4. Time
5. User Name
6. Pass Word
7. Log in

جدول ۱: پیش‌بینی مسیر حرکت ۲۰ کاربر مختلف وب سایت دانشگاه تربیت مدرس با استفاده از روش پیشنهادی و روش GIUDICI و مقایسه دقت جواب‌ها با هم.

شماره کاربر	ip کاربر	جواب حاصل از مدل پیشنهادی (x_1^*)		جواب حاصل از روش Giudici (x_2^*)	احتمال وقوع x_1^*	احتمال وقوع x_2^*
		جواب	نوع			
۱	۲۱۳.۱۷۶.۷۸.۱۰	$x_{.1}^* = x_{1.}^* = ۱$	G	$x_{.1}^* = x_{15}^* = x_{5.}^* = ۱$	۰/۰۸۴	۰/۰۴۲
۲	۶۵.۵۴.۱۸۸.۸۰	$x_{.9}^* = x_{9.}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۲۰	–
۳	۲۱۳.۱۷۶.۲۸.۶	$x_{.7}^* = x_{7.}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۲۸	–
۴	۱۹۴.۲۲۵.۶۲.۷۴	$x_{.1}^* = x_{15}^* = x_{5.}^* = ۱$	G	$x_{.1}^* = x_{15}^* = x_{56}^* = x_{6.}^* = ۱$	۰/۰۱۵	۰/۰۰۵۵
۵	۲۰۹.۲۳۷.۲۳۸.۱۷۹	$x_{.5}^* = x_{57}^* = x_{7.}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۱۱	–
۶	۲۱۷.۲۱۸.۳۶.۱۱۱	$x_{.2}^* = x_{2.}^* = ۱$	G	$x_{.6}^* = x_{63}^* = x_{2(10)}^* = x_{(10).}^* = ۱$	۰/۰۳۳	۰/۰۰۱
۷	۲۱۳.۱۷۶.۱۲۶.۷۴	$x_{.8}^* = x_{8.}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۱۵	–
۸	۲۴.۲۲.۲۳۲.۳۸	$x_{.6}^* = x_{67}^* = x_{7.}^* = ۱$	G	$x_{.6}^* = x_{6(10)}^* = x_{(10)1}^* = x_{14}^* = x_{4.}^* = ۱$	۰/۰۱۶	۰/۰۰۲
۹	۶۳.۱۴۸.۹۹.۲۳۹	$x_{.5}^* = x_{51}^* = x_{16}^* = x_{6.}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۰۵	–
۱۰	۶۲.۶۰.۱۹۶.۳	$x_{.1}^* = x_{15}^* = x_{53}^* = x_{33}^* = x_{3.}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۰۴	–
۱۱	۲۱۳.۱۷۶.۱۲۶.۵۹	$x_{.4}^* = x_{4.}^* = ۱$	G	$x_{.4}^* = x_{4.}^* = ۱$	۰/۰۲۸	۰/۰۲۸
۱۲	۸۲.۴۱.۱۴.۲۳۶	$x_{.1}^* = x_{15}^* = x_{53}^* = x_{33}^* = x_{3.}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۰۵	–
۱۳	۱۳۴.۲.۴۳.۱۱۹	$x_{.9}^* = x_{9.}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۲۱	–
۱۴	۲۱۳.۲۰۸.۴۵.۵	$x_{.1}^* = x_{15}^* = x_{5(10)}^* = x_{(10)9}^* = x_{9.}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۱۱	–
۱۵	۸۰.۱۹۱.۵۷.۱۳۸	$x_{.1}^* = x_{15}^* = x_{53}^* = x_{33}^* = x_{2(10)}^* = x_{(10).}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۰۳	–
۱۶	۱۹۴.۲۲۵.۱۶۶.۲۳۲	$x_{.1}^* = x_{15}^* = x_{54}^* = x_{42}^* = x_{2(10)}^* = x_{(10).}^* = ۱$	G	$x_{.1}^* = x_{15}^* = x_{54}^* = x_{42}^* = x_{2(10)}^* = x_{(10)3}^* = x_{36}^* = x_{6.}^* = ۱$	۰/۰۰۳	۰/۰۰۲۸
۱۷	۲۱۷.۲۱۹.۱۵۸.۱۵۵	$x_{.1}^* = x_{15}^* = x_{5.}^* = ۱$	G	$x_{.1}^* = x_{15}^* = x_{5.}^* = ۱$	۰/۰۳۳	۰/۰۳۳
۱۸	۲۱۳.۱۷۶.۲۸.۹	$x_{.6}^* = x_{67}^* = x_{7.}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۲۰	–
۱۹	۸۱.۹۱.۱۴۴.۲۵۰	$x_{.1}^* = x_{15}^* = x_{54}^* = x_{42}^* = x_{2(10)}^* = x_{(10).}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۰۳	–
۲۰	۲۱۷.۲۱۹.۱.۷	$x_{.1}^* = x_{15}^* = x_{5.}^* = ۱$	G	الگوریتم به لوپ می‌افتد	۰/۰۰۳	–

با توجه به این واقعیت که رفتار بازدید مشتریان از یک وب سایت فرآیندی پویاست، به‌منظور دست‌یابی به پیش‌بینی‌های دقیق‌تر لازم است تا ماتریس انتقال مدل مارکوفی هر کاربر به‌طور متناوب به روز شده تا به خوبی در برگیرنده آخرین تغییرات در سلایق و الگوهای بازدید مشتری مذکور باشد.

برای بهبود هرچه بیشتر عملکرد مدل ارائه‌شده و به‌منظور ارائه جوابی کامل‌تر می‌توان زمینه‌های تحقیقاتی آتی زیر را پیشنهاد نمود:

۱) واردکردن جنبه‌های دیگر بازدید (مثلاً مدت زمان بازدید یا مدت زمان بازدید هر صفحه) در مدل.

۲) استفاده از مدل‌های احتمالی دیگر به‌صورت ترکیبی برای یافتن جواب‌های دقیق‌تر.

۳) به‌کارگیری مدل‌های مارکوفی با درجات بالاتر.

۴) با توجه به محدودیت حافظه هر کاربر در نگهداری محتویات صفحات قبلی مشاهده‌شده به‌وسیله او، ظرفیت حافظه هر فرد را برای نگهداری اطلاعات مربوط به مشخصات صفحات مشاهده‌شده به‌وسیله وی محاسبه کرده و برای هر فرد جداگانه مدل مارکوفی با درجه‌ای برابر عدد ظرفیت حافظه وی استفاده شود.

سپاس‌گزاری

در انتها از پشتیبانی‌های بی‌دریغ مسئولین و کارکنان سایت مرکزی دانشگاه تربیت مدرس که با در اختیار نهادن داده‌های لازم برای اجرای این مدل، نویسندگان را در راستای پیشبرد اهداف این تحقیق یاری نمودند، تشکر و قدردانی می‌نماییم.

مراجع

[1] G. P. Shapiro, *Machine Learning and Data Mining*, Course Notes, 2004.

[2] A. Bestavros, "Speculative data dissemination and service to reduce server load, network traffic, and service time in distributed information system," in *Proc. of the 1996 Conf. on Data Engineering*, pp. 18 -187, 26 Feb.-1 Mar. 1996.

[3] I. Zuckerman, D. W. Albrecht, and A. Nicholson, "Predicting user's request on the WWW," in *Proc. of the 7th Int. Conf. on User Modeling*, pp. 275-284, Banff, Canada, 20-24 Jun. 1999.

[4] B. Huberman, P. Pirolli, J. Pitkow, and R. Lukose, "Strong regularities in World Wide Web surfing," *Science*, vol. 280. no. 5360, pp. 95-97, 3 April 1998.

[5] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, *Visualization of Navigation Patterns on a Web Site Using Model Based Clustering*, Technical Report, MSR-TR-00-18, Microsoft Research, Redmond, WA, US, 2000.

[6] V. A. Petrushin, "eShopper Modeling and Simulation," in *Proc. SPIE 2000 Conf. on Data Mining*, pp. 75-83, Beijing, China, 16-20 Oct. 2000.

[7] C. Theusinger and K. P. Huber, "Analysing the footsteps of your customers," 2000.

[8] A. Goldfarb, *Analysing Website Choice Using Clickstream Data*, Working Paper, Northwestern University, 2001.

[9] R. E. Bucklin and C. Sismeiro, "A model of web site browsing behavior estimated on clickstream data," *Journal of Marketing Research*, vol. 40, no. 3, pp. 249-67, 2001.

[10] W. W. Moe and P. S. Fader, "Capturing evolving visit behavior in clickstream data," *Journal of Interactive Marketing*, vol. 30, no. 1, pp. 5-19, Winter 2004.

[11] A. L. Montgomery, and S. Li, K. Srinivasan, and J. Liechty, *Predicting Online Purchase Conversion Using Web Path Analysis*, GSIA Working Paper, Nov. 2002.

[12] A. L. Montgomery and F. Christos, *Using Clickstream Data to Identify World Wide Web Browsing Trends*, GSIA Working Paper #2000-E20 2002.

[13] S. Li, J. Liechty, and A. Montgomery, *Modeling Category Viewership of Web Users with Multivariate Count Models*, GSIA Working Paper #2003-E25, 2002.
[14] P. Giudici and C. Tarantina, *Applied Data Mining*, Wiley, London, 2003.
[15] Y. Park and P. S. Fader, "Modeling browsing behavior at multiple websites," *Marketing Science*, vol. 23, no. 3, pp. 280-303, Summer 2004.
[16] A. L. Montgomery, S. Li, K. Srinivasan, and J. Liechty, "Modeling online browsing and path analysis using clickstream data," *Marketing Science*, vol. 23, no. 4, pp. 579-595, Fall 2004.
[17] P. Giudici and C. Tarantina, *Web Mining Pattern Discovery*, Technical Report #156, University of Pavia, Italy, Sep. 2003.
[18] E. Balas, "The prize collecting traveling salesman problem," *Networks*, vol. 19, pp. 621-636, 1989.
[19] E. Balas and G. Martin, "Roll-a-round: software package for scheduling the sounds of a rolling mill," *Balas and Martin Associates*, 1985.
[20] M. Fischetti and P. Toth, "An additive approach for the optimal solution of prize-collecting traveling salesman problem," in B. L. Golden and A. A. Assad (eds.), *Vehicle Routing: Methods and Studies*, North Holland, Amsterdam, 1988.
[21] E. H. A. Aarts, J. H. M. Korst, and P. J. M. Van Laarhoven, "A quantitative analysis of the simulated annealing algorithm: a case study for the traveling salesman problem," *J. Statistical Physics*, vol. 50, no. 1-2, pp. 187-206, 1988.
[22] J. R. A. Allwright and D. B. Carpenter, "A distributed implementation of simulated annealing for traveling salesman problem," *Parallel Computing*, vol. 10, no. 3, pp. 335-338, May 1989.
[23] D. Johnson and A. McGeoch, "The traveling salesman problem: a case study in local optimization," in E. H. L. Aarts and J. Lenstra (eds.), *Local Search in Combinatorial Optimization*, John Wiley & Sons, Inc., New York, 1995.

محمد مهدی سپهری تحصیلات خودرا در مقطع کارشناسی بازرگانی، گرایش بازاریابی در سال ۱۳۵۵ از مدرسه عالی بازرگانی تهران و در مقاطع کارشناسی ارشد و دکتری تحقیق در عملیات به ترتیب در سال‌های ۱۳۶۶ و ۱۳۷۰ از دانشگاه تنسی، ناکسویل آمریکا به پایان رسانده است. دکتر سپهری از سال ۱۳۷۰ در دانشکده فنی و مهندسی دانشگاه بوعلی سینا، همدان، و از سال ۱۳۷۵ در دانشگاه تربیت مدرس، تهران مشغول به فعالیت گردید و هم اکنون دانشیار بخش مهندسی صنایع دانشگاه تربیت مدرس است. زمینه‌های علمی مورد علاقه وی متنوع بوده و در بر گیرنده موضوعاتی مانند برنامه‌ریزی ریاضی و بهینه‌یابی در فناوری اطلاعات، بهینه‌یابی در پزشکی و درمان، داده‌کاوی و تجزیه‌وتحلیل متن، و تحلیل شبکه‌های اجتماعی است.

فؤاد مهدوی پژوه در سال ۱۳۸۲ مدرک کارشناسی مهندسی صنایع خود را از دانشگاه صنعتی شریف و در سال ۱۳۸۵ مدرک کارشناسی ارشد مهندسی صنایع خود را از دانشگاه تربیت مدرس دریافت نمود. وی از سال ۱۳۸۶ تاکنون مشغول به تحصیل در مقطع دکتری مهندسی صنایع در دانشگاه ایالتی اوکلاهمای آمریکا بوده و زمینه‌های تحقیقاتی مورد علاقه ایشان شامل برنامه‌ریزی ریاضی، بهینه‌سازی ترکیبیاتی و داده‌کاوی بر پایه گراف می‌باشد.