ارائه مدلی برای تحلیل و پیش‌بینی رفتار کاربران الکترونیکی مبتنی‌بر تکنیک‌های وب‌کاوی

انیس فرشیان عباسی[۱]، محمد فتحیان[۲*]، ابراهیم تیموری[۳]

۱. دانشجوی کارشناسی ارشد گروه مهندسی سیستم، دانشکده مهندسی صنایع، دانشگاه علم و صنعت، تهران، ایران
۲. استاد گروه مهندسی سیستم، دانشکده مهندسی صنایع، دانشگاه علم وصنعت ایران، تهران، ایران
۳. دانشیار گروه مهندسی سیستم، دانشکده مهندسی صنایع، دانشگاه علم و صنعت، تهران، ایران

دریافت: ۱۳۹۵/۱۲/۲۱ پذیرش: ۱۳۹۶/۰۴/۲۶

چکیده

امروزه ظهور خدمات مبتنی‌بر وب مانند تجارت الکترونیکی، بانک‌داری الکترونیکی و موارد مشابه موجب ایجاد تغییرات اساسی در روش زندگی انسان‌ها شده است. وب یک رسانه ارتباط مستقیم با هزینه کم را برای ارائه خدمات کسب و کارها به مشتریانشان فراهم می‌کند. کسب و کارها برای فعالیت‌های ترویجی و بازاریابی هدفمند نیاز به ثبت، بررسی و تحلیل رفتار کاربران و کشف دانش نهفته در آن را دارند تا بتوانند محتوا و ظاهر وب سایت خود را با علایق و نیازهای کاربران سازگار و شخصی‌سازی کنند. در این راستا برای تحلیل رفتار کاربران و ارائه پیشنهادات پویا و متناسب با الگوهای رفتاری آن‌ها می‌توان از تکنیک‌های وب‌کاوی استفاده کرد. در این پژوهش مدلی ارائه شده است که به کمک آن می‌توان رفتار کاربران الکترونیکی را تحلیل و پیش‌بینی کرد. در این مدل ابتدا کاربران به کمک الگوریتم انتشار کشش[۱] خوشه‌بندی شده‌اند و سپس به وسیله الگوریتم کاوش الگوهای ترتیبی سی.ام. اسپید[۲] رفتارشان تحلیل شده است. در گام بعد برای هر خوشه پروفایل کاربری مختص آن خوشه تشکیل می‌شود. سپس به کمک این پروفایل‌ها می‌توان توصیه‌هایی را به کاربران جدید ارائه کرد. نتایج به‌دست‌آمده حاکی از این است که مدل ارائه شده کارایی قابل قبولی دارد.

کلیدواژگان: وب‌کاوی، شخصی‌سازی، کاربردکاوی وب، خوشه‌بندی، کاوش الگوهای ترتیبی

---

*نویسندهٔ مسؤول مقاله: E-mail: fathian@iust.ac.ir

۱- مقدمه

شبکه جهانی وب[۳] حاوی حجم عظیمی از داده‌هاست و به مرور زمان این داده‌ها از لحاظ اندازه و حجم و هم میزان کاربردشان به طور نمایی رشد کرده و تغییر می‌کنند. این پدیده که گران‌بار شدن اطلاعات نامیده می‌شود، مشکلاتی را برای کاربران وب به وجود آورده است.

عدم دسترسی آسان به اطلاعات مورد نیاز از مهم‌ترین این مشکلات است. کاربران در این انباره عظیم برای یافتن اطلاعات مورد نیاز خود در زمان مناسب و به صورت آسان دچار مشکل هستند، زیرا از یک سو باید میزان ارتباط هر صفحه را با نیاز خود بررسی و از سوی دیگر باید صفحات را از نظر میزان قابلیت اعتماد ارزیابی کنند. با وجود سایت‌های رقیبی که تنها یک کلیک از وب سایت مورد نظر فاصله دارند، نیاز به افزودن قابلیت‌های اضافی به خدمات وب به عنوان لازمه ایجاد مشتری وفادار به وضوح احساس می‌شود. این قابلیت‌های اضافی تنها با تمرکز بر نیازها و علایق فردی مشتریان و فراهم کردن خدمات و محصولات متناسب با آن‌ها امکان‌پذیر است. در این راستا و برای انطباق محتوا و خدمات وب سایت با نیازمندی‌ها و علاقه‌مندی‌های کاربران می‌توان از تکنیک‌های وب‌کاوی استفاده کرد. وب‌کاوی از تکنیک‌های مختلف داده‌کاوی جهت استخراج دانش و اطلاعات جذاب از داده‌های موجود در وب استفاده می‌کند و اطلاعات مورد نیاز کاربران را برای آن‌ها فراهم می‌کند [۱، ص ٤۰؛ ۲، صص ۱۳-١٤؛ ۳، ص ۲۰]. در این پژوهش تلاش شده تا به کمک روش‌های نوین و بهینه وب‌کاوی رفتار کاربران الکترونیکی یک وب سایت نمونه ایرانی و فعال در حوزه تجارت الکترونیکی تحلیل و الگوهای رفتاری آن‌ها استخراج و رفتار پیمایشی آن‌ها پیش‌بینی شود.

۲- مروری بر مفاهیم اولیه

در دهه‌های اخیر توانایی بشر برای تولید و گردآوری داده‌ها به سرعت افزایش پیدا کرده است. این رشد انفجاری در داده‌های ذخیره شده سبب پیدایش فناوری جدیدی شده تا این حجم داده را به اطلاعات و دانش تبدیل کند. داده‌کاوی به عنوان یک راه‌حل برای این مسائل مطرح است [٤، ص ٦٦]. داده‌کاوی، استخراج یا اقتباس دانش از مجموعه داده‌هاست و به فرایندی گفته می‌شود که دانش را از داده‌ها استخراج می‌کند. این دانش در قالب الگوها و

مدل‌ها بیان می‌شود [٥، ص ١٦٤]. وب‌کاوی به کارگیری تکنیک‌های داده‌کاوی جهت کشف و استخراج خودکار الگوها و اطلاعات از داده‌های مربوط به دسترسی کاربران، ساختار ابرلینک‌های وب و محتوای صفحات وب است. وب‌کاوی صرفاً یک کاربرد خاص از داده‌کاوی نیست؛ زیرا داده‌های موجود در وب ماهیتاً ناهمگن، نیمه‌ساخت‌یافته یا غیرساخت‌یافته هستند. وب‌کاوی به دلیل پرمایگی و تنوع اطلاعات موجود در وب و سایر خصوصیات منحصر به فرد وب بسیاری از الگوریتم‌ها و تکنیک‌های منحصر به فرد خود را توسعه داده است. به طور کلی وب‌کاوی به سه دسته زیر تقسیم می‌شود.

۱- محتواکاوی وب[4]: این مفهوم به کشف اطلاعات مفید از محتوای صفحات وب (متن، داده‌های چندرسانه‌ای مانند تصاویر، ویدیو، صدا و...) اشاره دارد.

۲- ساختارکاوی وب[5]: به تحلیل، کشف و مدل‌سازی ساختار ارتباطات و اتصالات[6] صفحات وب اشاره دارد که برای ایجاد خلاصه وضعیت ساختار[7] استفاده می‌شود.

۳- کاربردکاوی وب: به درک رفتار کاربرانی که با وب سایت در تعامل هستند اشاره دارد. بدین منظور از لاگ فایل‌های[8] مختلف جهت استخراج دانش استفاده می‌کند. از این اطلاعات استخراج شده می‌توان برای سازمان‌دهی مجدد وب سایت، بهبود شخصی‌سازی و توصیه، بهبود لینک‌ها، پیمایش‌ها، جابه‌جایی‌ها و جذب تبلیغات بیشتر استفاده کرد. در نتیجه کاربران بیشتری جذب وب سایت می‌شوند و همین امر موجب افزایش درآمد می‌شود [۱، ص ٤٠]. شخصی‌سازی تعامل با کاربران نیز از کاربردهای مهم وب‌کاوی است.

شخصی‌سازی وب هر گونه اقدامی است که اطلاعات یا خدمات ارائه شده توسط یک وب سایت را با نیازهای یک کاربر یا گروه خاصی از کاربران با به کارگیری دانش به‌دست‌آمده از رفتار گردشی کاربر و علایق خاص او به صورت ترکیب با محتوا و ساختار وب سایت سازگار می‌کند. هدف یک سیستم شخصی‌سازی وب عبارت از فراهم کردن اطلاعات دلخواه یا مورد نیاز کاربران بدون درخواست صریح از آن‌هاست. هدف شخصی‌سازی وب براساس کاربردکاوی وب، توصیه کردن یک مجموعه از اشیا به کاربر جاری شامل لینک، آگهی، متن، محصول و غیره با جهت‌گیری به سمت ترجیحات و علایق کاربر است. این عمل با تطابق نشست[9] جاری کاربر با الگوهای کاربردی کشف شده از طریق کاربردکاوی وب صورت می‌گیرد (نشست کاربر، تمام صفحات درخواستی است که در یک بار بازدید سایت و در یک

بازه زمانی مشخص، کاربر آن‌ها را درخواست داده است). این فرآیند توسط موتور توصیه انجام می‌شود که مؤلفه برخط سیستم شخصی‌سازی است. فرآیند کلی شخصی‌سازی وب براساس کاربردکاوی وب شامل سه مرحله است: ۱- آماده‌سازی و پیش‌پردازش داده[۱۰]؛ ۲- کشف الگو[۱۱] (کشف دانش[۱۲])؛ ۳- تحلیل الگو[۱۳] و استفاده از الگوهای کشف شده برای شخصی‌سازی وب (تحلیل و ارائه دانش[۱٤])[۲، صص ۱٤-۲٥].

۳- پیشینه پژوهش

با توجه به مفاهیم قسمت پیشین در زمینه وب‌کاوی و شخصی‌سازی، اهمیت به کارگیری وب‌کاوی در تحلیل رفتار کاربران و فراهم آوردن محتوا و محیط شخصی‌سازی شده مناسب برای آن‌ها مشخص است. در ادامه یک سری از مقالات حوزه وب‌کاوی و تحلیل رفتار کاربران مورد بررسی قرار گرفته است. در یک پژوهش جهت شناسایی نشست‌های کاربران، الگوریتم ابتکاری مبتنی‌بر وقفه[۱٥] پیشنهاد شده است. به‌کارگیری الگوریتم‌های خوشه‌بندی مبتنی‌بر چگالی مانند دی.بی.اسکن[۱٦] برای کشف الگوهای جابه‌جایی پیشنهاد شده است که کاربران با استفاده از این الگوریتم‌ها خوشه‌بندی می‌شوند [٦، ص ٥٠]. در پژوهش بعدی نیز خوشه‌های کاربران و خوشه‌های صفحات وب توسط دو روش تحلیل بردار[۱۷] و نظریه فازی[۱۸] مشخص می‌شوند. در روش به‌کاربرده شده در این مقاله نیازی به شناسایی نشست‌های کاربر از لاگ‌های وب سرور نیست، همچنین الگوریتم شناسایی مسیرهای دستیابی پرتکرار ارائه شده در این پژوهش مبتنی‌بر کاوش الگوهای ترتیبی نیست [۷، ص ٦۲۲].

در مقاله‌ای دیگر روشی ارائه شده است که محتواکاوی و کاربردکاوی وب در آن با هم تلفیق شده‌اند. در این مقاله پس از پیش‌پردازش متون از الگوریتم سی.اف.دابلیو.اس[۱۹] برای خوشه‌بندی متون، و برای خوشه‌بندی کاربران از الگوریتم کی- مینز[۲۰] استفاده شده است. جهت کاوش قواعد باهم‌آیی و به منظور کاوش مجموعه‌های متناوب و مکرر نیز الگوریتم اپریوری[۲۱] به‌کار برده شده است [۸، صص ٥٤-٦٤]. برخی از محققین برای کشف الگوهای جابه‌جایی روشی بر مبنای تکنیک دنباله‌کاوی[۲۲] پیشنهاد کرده‌اند و بدین منظور دو کار اصلی انجام داده‌اند؛ نخست استفاده از گراف ردپا[۲۳] جهت مصورسازی داده‌های رشته کلیکی کاربر که موجب سهولت و سرعت در شناسایی الگوهای جذاب می‌شود، دوم ارائه روش

دنباله‌کاوی جهت شناسایی خودکار الگوهای جابه‌جایی کاربران با استفاده از رفتارهای جابه‌جایی برخط آن‌ها و نیز تلفیق آن با مدل شبکه پس‌انتشار[۲۴] جهت پیش‌بینی نیازهای بالقوه آتی کاربران است [۹، ص ۲۸۹۸].

در پژوهشی دیگر جهت استخراج اطلاعات مفید از داده‌های کاربرد وب از تکنیک خوشه‌بندی و الگوریتم کی- مینز استفاده شده است. همچنین قواعد باهم‌آیی و الگوریتم اپریوری جهت استخراج روابط جالب میان اعضای خوشه‌ها پس از خوشه‌بندی استفاده شده است. الگوریتم NMEEF-SD نیز به‌کار گرفته شده که این الگوریتم سیستم فازی تکاملی است که هدف آن استخراج قواعد فازی توصیفی و انجام عملیات کشف زیرگروه‌هاست [۱۰، صص ۱۱۲۴۳-۱۱۲٤۸].

در مطالعه دیگری تکنیک‌های کاربردکاوی وب و محتواکاوی با هم ترکیب شده و سه هدف اصلی زیر را دنبال می‌کند؛ نخست ایجاد پروفایل‌های جابه‌جایی کاربران[۲۵] جهت استفاده برای پیش‌بینی لینک‌ها[۲٦] به کمک الگوریتم‌های پی.ای.ام[۲۷] و اسپید[۲۸]، دوم غنی کردن پروفایل‌ها با اطلاعات معنایی[۲۹] به منظور تنوع بخشیدن به آن‌ها به کمک STMT[۳۰]، سوم به‌دست‌آوردن مشتریان جهانی با زبان‌های مختلف که اطلاعات مهمی را برای طراحی‌های آینده وب سایت دراختیار می‌گذارد [۱۱، صص ۷٤۷۸-۷٤۸۹]. در مقاله‌ای دیگر برای تحلیل رفتار کاربران و ارائه پیشنهاد به آن‌ها در ابتدا ترجیحات کاربران به کمک روش‌های کاربردکاوی وب به صورت خودکار از رشته کلیکی استخراج می‌شود؛ یعنی تنها رکوردهای خرید کاربر را در نظر نمی‌گیرد و کل رشته کلیکی او را در نظر می‌گیرد. در مرحله بعد به کمک درخت تصمیم مشتریانی انتخاب می‌شوند که احتمال بیشتری دارد که محصولات پیشنهادی را بخرند. در مرحله بعد با استفاده از قواعد باهم‌آیی محصولات مناسب‌تر برای پیشنهاد انتخاب می‌شوند [۱۲، ص ۳۲۹].

در پژوهشی دیگر روش خوشه‌بندی مبتنی‌بر دنباله توسط پیشنهاد دنباله جدید مرتبط با مدل مارکوفی توسعه داده شده است. الگوریتم خوشه‌بندی کی- مینز توسعه‌یافته با ART نیز ارائه شده است [۱۳، ص ۵۱۲].

در مطالعه‌ای دیگر برای پیش‌بینی الگوهای کاربردی از الگوریتم مبتنی‌بر کلونی مورچگان استفاده شده است. در این پژوهش سه منبع داده مرتبط با وب‌کاوی یعنی محتوا، ساختار و

کاربرد وب مورد استفاده قرار گرفته‌اند [۱٤، ص۸۸۹]. پژوهشگران دیگری به منظور خوشه‌بندی کاربران از الگوریتم کی- مینز استفاده کرده‌اند. سپس داده‌های این خوشه‌ها با توجه به سن، جنسیت، زمان برخط[۳۱] بودن، آدرس صفحه بازدید شده، زبان و رفتار تقسیم‌بندی شده‌اند [۱۵، صص ۱-٤].

در مقاله‌ای دیگر رفتار افراد مسنی که از یک نرم‌افزار کاربردی مربوط به حوزه پایش سلامت و مراقبت از خود[۳۲] استفاده می‌کنند، توسط تکنیک‌های وب‌کاوی تحلیل شده است. ابتدا دوره استفاده از این نرم‌افزار، زمان، تعداد عملیات، میزان و حجم خدمات ارائه شده توسط این نرم‌افزار به کاربران مختلف گردآوری و تحلیل شده، سپس از الگوریتم‌های کاوش قواعد باهم‌آیی برای کشف ارتباط میان این عوامل استفاده شده است. در مرحله بعد، نشست‌های کاربران ایجاد می‌شود و توسط نوع خاصی از الگوریتم خوشه‌بندی کی- مینز (کی- مینز توسعه‌یافته با ART2[۳۳])، این داده‌ها خوشه‌بندی می‌شوند. در انتها نیز پروفایل‌های ترتیبی مرتبط با الگوهای رفتاری کاربران به کمک اعمال شمای مبتنی‌بر دنباله[۳٤] همراه با مدل‌های مارکوفی و الگوریتم کی- مینز توسعه‌یافته با ART2 برای کاوش رفتارهای دنباله‌ای کاربران خوشه استخراج شده‌اند [۱٦، ص۷۷۵].

ترکیب کاربردکاوی وب و محتواکاوی وب در پژوهشی دیگر نیز مورد استفاده قرار گرفته است. کاربران در مرحله کاربردکاوی توسط الگوریتم کی- مینز خوشه‌بندی شده‌اند. در مرحله بعد و برای محتواکاوی محتوای متنی صفحات وب به وسیله استخراج N-گرم‌های کاراکتری به‌دست می‌آیند. نتایج به‌دست‌آمده از مرحله کاربردکاوی و محتوا کاوی با یکدیگر تلفیق شده و جهت تحلیل و ارائه توصیه از این نتایج استفاده می‌شود [۱۷، صص ۱-۳].

برخی دیگر از پژوهشگران در پژوهش خود بر ایجاد سیستم توصیه‌گر بی‌درنگ و پویا برای تمام بازدیدکنندگان یک وب سایت (صرف‌نظر از این‌که بازدیدکنندگان در سایت ثبت‌نام کرده یا نکرده‌اند) تمرکز دارند. در این تحقیق سعی شده است تا پیشنهادات مؤثر و کارا به تمام بازدیدکنندگان سایت ارائه شود. سیستم پیشنهادی توانسته بر یکسری از محدودیت‌های سیستم‌های توصیه‌گر رایج و سنتی غلبه کند [۱۸، صص ٦۰-٦٤].

با توجه به بررسی‌های انجام شده و نتایج حاصل از مرور پیشینه پژوهش می‌توان دریافت که یکی از شکاف‌های تحقیقاتی اساسی در این حوزه عدم به‌کارگیری الگوریتم‌های

خوشه‌بندی غیروابسته به تعداد خوشه‌ها (k) و بررسی میزان کارایی آن‌ها در این حوزه است. به همین دلیل در این پژوهش از الگوریتمی که به k وابسته نیست استفاده شده و پس از آن از یکی از الگوریتم‌های بهینه کاوش الگوهای ترتیبی برای تحلیل رفتارهای کاربران هر خوشه بهره گرفته شده است. در انتها نیز با توجه به رفتار کاربران پیشنهاداتی به آن‌ها ارائه شده است.

۴- مدل پیشنهادی

با توجه به پیشینه پژوهش (به ویژه تحقیقات انجام شده در مراجع ۱۱، ۲۰، ۲۱) و به منظور رفع شکاف تحقیق که در قسمت پیشین بیان شد، مدل پیشنهادی در ادامه ارائه می‌شود. این مدل به منظور تحلیل رفتار کاربران وب استفاده شده است. نخست داده‌های لاگ فایل پیش‌پردازش می‌شوند. سپس الگوریتم خوشه‌بندی انتشار کشش که به مقدار k (تعداد خوشه‌های کاربر) وابسته نیست، بر آن‌ها اعمال می‌شود. در مرحله بعد روی داده‌های هر خوشه الگوریتم کاوش الگوهای ترتیبی سی.ام. اسپید[۳۵] اعمال می‌شود تا الگوهای پرتکرار مربوط به هر خوشه استخراج شوند. در گام نهایی نیز داده‌ها برای اعتبارسنجی مدل ارائه شده به دو دسته داده آموزشی و آزمایشی تقسیم می‌شوند و به کمک صفحات پرتکرار در هر خوشه و نیز الگوریتم دسته‌بندی k همسایه نزدیک[۳۶] معیارهایی از قبیل دقت و فراخوانی محاسبه می‌شود، در نهایت کارایی این روش مشخص می‌شود. در شکل ۱ نمایی از مدل ارائه شده جهت تحلیل رفتار کاربر مشاهده می‌شود که در آن از سه روش کاربردکاوی وب (خوشه‌بندی، دسته‌بندی، کاوش الگوهای ترتیبی) استفاده شده است.

۴-۱- مرحله پیش‌پردازش

داده‌هایی که در این پژوهش مورد استفاده قرار گرفته لاگ دسترسی مربوط به وب سایت یک انتشارات معروف و با سابقه ایرانی است که در زمینه تجارت الکترونیکی و فروش محصولات و کتاب‌هایش بیشتر از ۵ سال است که به صورت الکترونیکی فعالیت دارد. داده‌های به کار برده شده در بازه زمانی ده روز، از تاریخ ۲۰۱٦/۰۸/۰۱ تا تاریخ ۲۰۱٦/۰۸/۱۰ گردآوری شده است (تقریباً اواسط مرداد ۱۳۹۵). لاگ دسترسی از نوع W3C

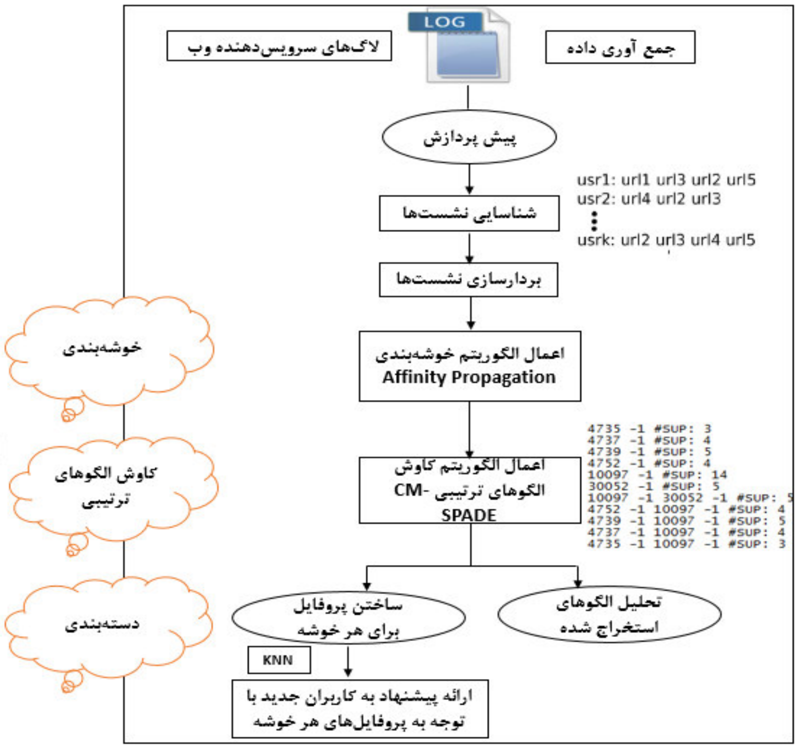
(یکی از انواع فرمت‌های ذخیره لاگ) است و فیلدهای اصلی این نوع لاگ را دارد. در مرحله پیش‌پردازش داده از نرم‌افزار متلب بهره گرفته شده است. مراحل پیش‌پردازش پس از بارگزاری داده‌ها در متلب به ترتیب زیر انجام می‌شوند.

- **پاک‌سازی داده و فیلترینگ**[۳۷]**:** در ابتدا فیلدهای مجموعه داده بررسی می‌شود و موارد اضافه و غیرضروری حذف می‌شوند. برای مثال فقط روش Get (method) در نظر گرفته شده است. رکوردهای مربوط به فایل‌های گرافیکی و اسکریپت‌ها ودرخواست‌های مربوط به مدیریت سایت نیز حذف می‌شوند، همچنین درخواست‌های مربوط به اخبار نیز به دلیل متغیر بودن حذف می‌شوند [۱۱، صص ۷۴۸۲-۷۴۸۳]. این گام فیلترینگ نامیده می‌شود.

- **انتخاب رکوردهای معتبر یا دی-اسپایدرینگ**[۳۸]: در این مرحله درخواست‌هایی که از طرف ربات‌ها[۳۹] فرستاده شده در لاگ فایل حذف می‌شوند. برای این کار فیلد عامل کاربر[۴۰] بررسی می‌شود.

-**ایجاد برچسب زمانی**: در این گام باید زمان پایه و شروع کار مشخص شود. برای نمونه ساعت ۰۰:۰۰:۰۰ روز ۰۱/ ۰۸/ ۲۰۱۶ به عنوان زمان پایه در نظر گرفته شده است. سپس در هر رکورد زمان به ثانیه محاسبه شده و تفاوتش با زمان پایه حساب شده و در نهایت در یک فیلد جداگانه ذخیره می‌شود.

- **تخصیص شماره صفحه منحصر به فرد به هر صفحه از وب سایت:** پس از این‌که رکوردهای غیرضروری حذف شدند و تنها رکوردهای مورد نیاز در مجموعه داده باقی ماند، باید مشخص شود که هر رکورد مربوط به کدام آدرس صفحه وب سایت است و سپس شماره منحصر به فرد به آن صفحه اختصاص داده شود. تخصیص شماره به هر آدرس با بررسی آدرس آن صفحه انجام شده و با توجه به اعدادی که در آدرس هر صفحه موجود است، شماره صفحه‌های منحصر به فرد به‌دست آمده است. در این گام نیز فیلدهای یو.آر.آی[۴۱] بررسی شده و الگوریتم ابتکاری با توجه به رشته موجود در آن و اعداد موجود در رشته از طریق فرمول‌های معین می‌تواند شماره صفحات را مشخص کند.



شکل ۱ مدل ارائه شده برای تحلیل و پیش‌بینی رفتار کاربران

– **شناسایی کاربران:** در این گام ویژگی (فیلد) آدرس آی‌پی[42] کاربران بررسی شده و به کمک آن و نیز ویژگی عامل(حاوی اطلاعاتی از مرورگر کاربر)، کاربران منحصر به فرد شناسایی شده‌اند. راهبرد واکنشی[43] راهبردی است که برای تشخیص کاربران و نشست آن‌ها در این پژوهش در نظر گرفته شده است.

– **تشخیص نشست کاربران:** در این مرحله به منظور شناسایی نشست‌های هر کاربر حد آستانه t برابر با عدد رایج ۳۰ دقیقه در نظر گرفته شده است (یعنی طول هر نشست بیش از سی دقیقه نیست). سپس برای هر کاربر تفاوت زمانی میان دو سطر یا دو رکورد لاگ فایل بررسی می‌شود. اگر تفاوت به‌دست‌آمده بیش از t باشد نشست جدیدی برای آن کاربر ایجاد

می‌شود. در غیر این صورت صفحه آن رکورد جزء نشست اخیر محسوب می‌شود.

مرحله پیش‌پردازش داده به علت گوناگونی زیاد داده‌ها بسیار زمان‌بر است و رکوردهای لاگ فایل باید تقریباً سطر به سطر بررسی می‌شدند و موارد حذفی و همچنین مخدوش مشخص می‌شدند. از طرفی تعداد صفحات سایت به نسبت زیاد و آدرس آن‌ها متغیر بود و استخراج الگو از میان این آدرس‌های متغیر به منظور تخصیص شماره صفحه زمان زیادی را گرفت. جدول ۱ اطلاعاتی درباره داده‌های به دست آمده پس از مراحل مختلف را در اختیار قرار می‌دهد.

جدول ۱ تعداد داده‌های به‌دست‌آمده در هر مرحله

<<عنوان>>	<<تعداد>>
کل رکوردهای لاگ فایل در بازه ده روز	۳۹۱۹۳۹
کل رکوردهای لاگ فایل پس از پیش‌پردازش	۱۴۷۱۲
نشست‌های شناسایی شده	۸۴۹
کل صفحات وب منحصربه فرد بازدید شده در بازه ده روز	۱۰۲۳
صفحات وب منحصربه فردی که بیش از ده بار در بازه ده روز به آن‌ها مراجعه شده است	۲۲۱
تعداد کاربران در بازه ده روز	۳۵۱۴
تعداد نشست‌های مجموعه آموزشی	۵۹۴
تعداد نشست‌های مجموعه آزمایشی	۲۵۵

۴-۲- کاربرد کاوی وب

پس از تشخیص نشست‌های کاربران، نخست بردارهای مربوط به نشست‌ها ساخته می‌شود و سپس نشست‌ها خوشه‌بندی و از یکدیگر تفکیک می‌شوند و در خوشه‌های مناسب قرار می‌گیرند، سپس به کمک الگوریتم کاوش الگوهای ترتیبی الگوهای پرتکرار در هر نشست استخراج می‌شوند.

- بردارسازی نشست

پس از مشخص شدن نشست‌ها برای این‌که بتوان الگوریتم خوشه‌بندی مورد نظر را روی نشست‌ها اعمال کرد باید آن‌ها را به بردار تبدیل کرد. هر نشست به صورت یک بردار مشخص

می‌شود. S_i نشست i–ام یک کاربر است که به صورت رابطه (۱) تعریف می‌شود [۱۷].

$$S_i = \langle w(P_1, S_i),\ w(P_2, S_i), \ldots,\ w(P_k, S_i), \ldots,\ w(P_n, S_i) \rangle \quad (۱)$$

n تعداد صفحات وبی است که در همه جلسات دسترسی کاربران بازدید شده‌اند. P_k صفحه k–ام است و $w(P_k, S_i)$ نشان‌دهنده وزن و عددی است که باید در خانه k–ام بردار مربوط به S_i قرار گیرد. این وزن با توجه به معیار بسامد محاسبه می‌شود. بسامد در واقع تعداد بازدید از یک صفحه وب است. فرض بر این است که صفحات با بسامد بالاتر محبوبیت بیشتری نزد کاربران دارند که این امر طبیعی به نظر می‌رسد. بسامد هر صفحه در هر نشست از طریق رابطه (۲) محاسبه می‌شود [۱۷].

$$\text{Frequency(page)} = \frac{\text{number of visits in the session(page)}}{\sum \text{number of visits in the session(page)}} \quad (۲)$$

صورت این کسر نشان‌دهنده تعداد بازدیدهای کاربر از یک صفحه در یک نشست مشخص است. مخرج آن بیانگر تعداد کل بازدیدها از صفحات وب در همان نشست مشخص است. در پایان نیز تمام بردارهای مربوط به نشست‌های دسترسی کاربران در کنار هم قرار گرفته و ماتریس m*n بعدی از وزن‌های صفحات وب تشکیل می‌شود (m تعداد کل نشست‌های کاربران). سطرهای این ماتریس نشان‌دهنده نشست‌های کاربران است و ستون‌های آن نشان‌دهنده صفحاتی است که در نشست‌های مختلف وب، بازدید شده‌اند. اگر تعداد صفحات یا n از اندازه معقولی تجاوز کند نه تنها زمان پردازش بسیار زیادی را به هنگام خوشه‌بندی نشست‌ها مصرف می‌کند بلکه کاربست‌پذیری سیستم در جهان واقعی را نیز محدود می‌کند [۱۷، صص ٤٤–٤٦] .به منظور کاهش ابعاد صفحاتی که کمتر از ۱۰ بار در همه نشست‌ها به آن‌ها مراجعه شده است، از ماتریس حذف می‌شوند. نشست‌هایی شامل دو صفحه و یا کمتر، یا شامل بیش از ۸٦ صفحه نیز از ماتریس حذف می‌شوند [۱۱، صص ۷٤۸۱–۷٤۸٤].

- الگوریتم خوشه‌بندی انتشار کشش

تکنیک‌های خوشه‌بندی از راه شناسایی گروه‌هایی از کاربران که به نظر می‌رسد اولویت‌های مشابهی داشته باشند و تقسیم‌بندی گروه‌هایی که اولویت‌های بسیار متفاوتی دارند، کار می‌کنند. به طور خلاصه یک روش خوشه‌بندی خوب خوشه‌هایی را با کیفیت بالا تولید می‌کند، به طوری که درون هر خوشه بیشترین شباهت و بین خوشه‌های متفاوت،

کمترین شباهت وجود داشته باشد [۱۹، صص۵۷-۵۸]. همان‌طور که در مقاله [۲۰، صص ۹۷۲-۹۷۳] توضیح داده شده الگوریتم انتشار کشش الگوریتم خوشه‌بندی بر مبنای مفهوم تبادل پیام[۴۴] میان نقاط داده است. برخلاف الگوریتم‌های خوشه‌بندی مانند کی- مینز و کا- مدویدز، در این روش نیازی به مشخص بودن تعداد خوشه‌ها پیش از شروع الگوریتم نیست و همین ویژگی یکی از مزیت‌های اصلی این الگوریتم است. این الگوریتم نیز مشابه روش کا- مدویدز، نمونه‌ها یا الگوهایی را از میان داده‌های ورودی و به عنوان نماینده[۴۵] خوشه‌ها انتخاب می‌کند. الگوریتم انتشار کشش هم‌زمان تمام نقاط داده را به عنوان نمونه‌های بالقوه در نظر می‌گیرد. به کمک نمایش هر داده به صورت یک گره در شبکه، این روش به طور بازگشتی پیام‌های حاوی مقادیر حقیقی[۴۶] را تا زمانی که مجموعه مناسب از نمونه‌ها و خوشه‌های متناظرشان به‌دست آید، از طریق یال‌ها در داخل شبکه ارسال و تبادل می‌کند.

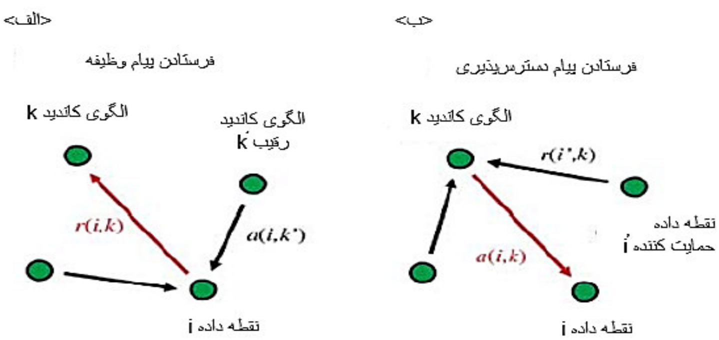
یکی از ورودی‌های الگوریتم انتشار کشش مجموعه‌ای از مشابهت‌هایی[۴۷] با مقادیر حقیقی میان نقاط داده است و S(i,k) بیانگر این است که چقدر داده با اندیس k جهت نمونه بودن و الگو بودن برای داده i مناسب است. اگر هدف کمینه کردن مربعات خطا باشد، هر مشابهتی با یک مقدار منفی مربع خطا (فاصله اقلیدسی) مقداردهی می‌شود (مطابق رابطه (۳)).

$$S(i,k)=-\|x_i - x_k\|^2 \leftarrow \text{برای نقاط } x_k \text{ و } x_i \qquad (۳)$$

ورودی دیگر این الگوریتم مقادیر ترجیح[۴۸] است. این الگوریتم به جای از پیش تعیین کردن تعداد خوشه‌ها مقدار حقیقی S(k,k) برای هر داده k را به عنوان ورودی در نظر می‌گیرد؛ به طوری که نقاطی با S(k,k) بزرگ‌تر احتمال بیشتری دارد که به عنوان نمونه و الگو انتخاب شوند. این مقادیر، مقادیر ترجیح یا خودمشابهت[۴۹] نامیده می‌شوند. در این تحقیق مقدار ترجیح یا خودمشابهت برای هر نقطه برابر با مقدار کمینه مشابهت‌های آن نقطه با سایر نقاط در نظر گرفته شده است(جهت توضیحات بیشتر به مرجع ۲۰ مراجعه شود).

دو نوع پیام می‌تواند میان نقاط داده تبادل شود که هر کدام نوع متفاوتی از رقابت را در نظر می‌گیرد: ۱- پیام وظیفه[۵۰]: r(i,k) که از نقطه i به نقطه نمونه کاندید k فرستاده می‌شود، بیانگر مجموع دلایلی است که نشان می‌دهد چقدر نقطه k، در مقایسه با سایر نقاط بالقوه نمونه، به عنوان نقطه نمونه برای نقطه i مناسب است. (شکل ۲-الف). ۲- پیام دسترس‌پذیری[۵۱]: a(i,k) که از نقطه نمونه کاندید k به نقطه i فرستاده می‌شود، بیانگر مجموع دلایلی است که نشان می‌دهد چقدر برای

نقطه i بهتر و مناسب‌تر است که نقطه k را به عنوان نمونه برای خودش انتخاب کند (شکل ۲-ب).



شکل ۲. پیام‌های وظیفه و دسترس‌پذیری در الگوریتم انتشار کشش [۲۰]

این پیام‌ها می‌توانند در هر مرحله‌ای با هم ترکیب شوند تا مشخص شود که کدام نقاط نمونه هستند و همچنین نقاط دیگر به کدام یک از نقاط نمونه تعلق دارند؛ بنابراین پس از محاسبه مشابهت‌های میان نقاط داده به عنوان ورودی الگوریتم هر تکرار از الگوریتم انتشار کشش شامل گام‌های زیر است: ۱- به‌روزرسانی تمام مقادیر وظیفه با توجه به مقادیر دسترس‌پذیری؛ ۲- به‌روزرسانی تمام مقادیر دسترس‌پذیری با توجه به مقادیر وظیفه؛ ۳- تلفیق مقادیر وظیفه و دسترس‌پذیری به منظور بررسی نتایج نمونه‌ها و خاتمه الگوریتم در صورتی که این نتایج پس از تعداد تکرار معینی تغییر نکنند.

در این پژوهش برای پیاده‌سازی این الگوریتم از نرم‌افزار متلب استفاده شده است. همان طور که به آن اشاره شد الگوریتم انتشار کشش نیز به یک‌سری ورودی نیاز دارد. مشابهت‌های[52] میان نقاط داده یکی از ورودی‌های این الگوریتم است. برای محاسبه مشابهت‌ها از منفی مربع خطا (فاصله اقلیدسی) در این الگوریتم استفاده می‌شود و به آسانی و با کمک دستورات ساده در متلب می‌توان مشابهت‌ها را برای تمام نقاط داده محاسبه کرد. مقادیر ترجیح، ورودی دیگری است که این الگوریتم نیاز دارد که در این تحقیق برابر با مقدار کمینه مشابهت‌ها در نظر گرفته شده است، سپس بردارهای به‌دست‌آمده از مرحله بردارسازی نشست به دو مجموعه آموزشی (٪۷۰ داده‌ها) و آزمایشی (٪۳۰

داده‌ها) تقسیم می‌شوند. پس از تنظیم مقادیر ورودی این الگوریتم روی داده‌های مجموعه آموزشی اعمال و خوشه‌ها مشخص می‌شوند. الگوریتم انتشار کشش، تعداد ١٤ خوشه را برای داده‌های آموزشی این پژوهش تشخیص داده است. در این الگوریتم مراکز و نمونه خوشه‌ها از میان داده‌ها انتخاب می‌شوند. نشست‌های شماره ٨، ١٣، ٤٦، ١٦٦، ١٨٦، ٢٠٨، ٢٤٢، ٣٠٢، ٣٠٤، ٣١٧، ٤٢٠، ٤٢١، ٥٥٤ و ٥٦٤ به عنوان نمونه این ١٤ خوشه انتخاب شده‌اند. الگوریتم انتشار کشش ثبات به نسبت خوبی دارد، حتی پس از چندین بار اجرا همان تعداد ١٤ خوشه را تشخیص می‌دهد.

- شناسایی الگوهای پرتکرار در هر خوشه به کمک الگوریتم سی.ام. اسپید

پس از این‌که داده‌ها خوشه‌بندی شدند جهت تحلیل رفتار کاربر از کاوش الگوهای ترتیبی استفاده شده است. الگوریتم کاوش الگوهای ترتیبی که در این پژوهش از آن بهره گرفته شده الگوریتم «سی.ام. اسپید» است. با توجه به [٢١، صص ٤٠-٤٢] الگوریتم سی.ام. اسپید کارایی و سرعت بیشتری در مقایسه با سایر الگوریتم‌های کاوش الگوهای ترتیبی پرکاربرد نظیر اسپید و GSP و Prefixspan دارد. به همین دلیل این الگوریتم برای به‌کارگیری انتخاب شده است. این الگوریتم بر مبنای الگوریتم اسپید است. اصلی‌ترین تفاوت میان سی.ام. اسپید و اسپید در این است که سی.ام. اسپید از روش جدیدی به نام هرس هم‌زمان یا با هم‌جهت[٥٣] در فضای جستجو استفاده می‌کند و همین امر موجب سرعت بیشتر این الگوریتم می‌شود.

الگوریتم اسپید از فرمت پایگاه داده عمودی[٥٤] استفاده می‌کند. به‌کارگیری چنین فرمتی این مزیت را دارد که می‌توان استخراج الگوها و محاسبه پشتیبانشان را بدون اجرای اسکن‌های پرهزینه پایگاه داده انجام داد. همین امر موجب می‌شود که الگوریتم‌هایی که از این فرمت استفاده می‌کنند کارایی[٥٥] بهتری نسبت به سایر الگوریتم‌ها داشته باشند که از فرمت پایگاه داده افقی[٥٦] استفاده می‌کنند، همچنین الگوریتم‌هایی که از فرمت پایگاه داده عمودی استفاده می‌کنند یک مشکل اساسی دارند و که تعداد زیادی الگوی کاندید را که در پایگاه داده ورودی وجود ندارند و یا در آن پرتکرار نیستند در مراحل خود تولید می‌کنند. به کمک روش هرسی که در الگوریتم سی.ام. اسپید استفاده شده است این مشکل تا اندازه بسیاری برطرف شده و زمان اجرا کاهش و کارایی بهبود یافته است. ورودی این الگوریتم پایگاه داده‌ای از دنباله‌ها[٥٧] و مقدار حداقل پشتیبان[٥٨] تعیین شده توسط کاربر است. یک پایگاه دنباله مجموعه‌ای از دنباله‌هاست که هر دنباله فهرستی از مجموعه اقلام است. یک مجموعه قلم[٥٩] مجموعه‌ای بدون ترتیب از اقلام[٦٠]

و عناصر است (فرض می‌شود که اقلام یک مجموعه قلم به ترتیب الفبا مرتب شده است). برای نمونه {a,b} یک مجموعه قلم، a یک قلم یا عنصر و b قلم یا عنصر دیگری است. ترتیب میان مجموعه اقلام مختلف در هر دنباله حائز اهمیت است و در واقع وجه تمایز اصلی بین کاوش قواعد باهم‌آیی[۶۱] و کاوش الگوهای ترتیبی در لحاظ کردن ترتیب زمانی است. در این پژوهش موارد زیر در نظر گرفته شده است: ۱- هر صفحه وب به عنوان یک قلم یا عنصر در نظر گرفته شده است؛ ۲- هر نشست به عنوان یک دنباله در نظر گرفته شده که فهرستی از مجموعه اقلام است، به صورتی که هر مجموعه قلم تنها شامل یک صفحه وب می‌شود.

برای به‌کارگیری این الگوریتم ازکتابخانه متن باز[۶۲] داده‌کاوی SPMF که با جاوا نوشته شده و در زمینه کاوش الگو نیز به طور اختصاصی توسعه یافته استفاده شده است. داده‌های هر یک از خوشه‌های شناسایی شده به فرمت خاصی تبدیل شده و به این الگوریتم وارد می‌شوند. سپس باید یک مقدار حداقل پشتیبان[۶۳] به عنوان ورودی این الگوریتم مشخص شود. مقداری که در این پژوهش با توجه به طبیعت داده‌ها و حجم بالای رکوردهای موجود در لاگ فایل به منظور استخراج تعداد مناسبی از الگوها در نظر گرفته شده ۰/۲ است. پس از اعمال این الگوریتم صفحات پرتکرار و نیز به طور کل الگوهای پرتکرار (دنباله صفحاتی که بیش از بقیه بازدید شده‌اند) در هر خوشه مشخص می‌شود. در انتها نیز با توجه به نتایج حاصل از اعمال این الگوریتم پروفایل کاربری برای هر خوشه تعیین می‌شود. پروفایل هر خوشه نشان‌دهنده صفحات پرکاربرد آن خوشه است [۲۲]. در این پژوهش در پروفایل هر خوشه تنها دو صفحه در نظر گرفته شده است.

۴-۴- تحلیل نتایج حاصل از اعمال الگوریتم سی.ام. اسپید

نتایج حاصل از اعمال الگوریتم کاوش الگوهای ترتیبی مانند سی.ام. اسپید می‌تواند درباره رفتار کاربران وب سایت اطلاعات مفیدی در اختیار مدیران و توسعه‌دهندگان آن سایت قرار دهد. تحلیل رفتار کاربران خوشه‌ها در جدول ۲ نشان داده شده است. به کمک نتایج حاصل از این تحلیل می‌توان دریافت که کاربران هر خوشه چه رفتاری از خود و به چه مواردی علاقه بیشتری نشان داده‌اند. سپس با بررسی این نتایج مدیران سایت می‌توانند تصمیم بگیرند که به کاربران خوشه‌های مختلف چه مواردی را پیشنهاد دهند و یا با توجه به خوشه کاربر تخفیف‌های مختلف

و پیشنهادهای جدید ارائه دهند و بدین طریق هم میزان رضایت کاربر را افزایش دهند و هم سود بیشتری کسب کنند.

جدول ۲ نتایج حاصل از تحلیل الگوهای به‌دست‌آمده برای هر خوشه

شماره خوشه	مرکز خوشه	نتایج حاصل از تحلیل
خوشه ۱	رکورد شماره ۸	کاربران این خوشه به کتاب‌هایی با موضوع ریاضیات، دیفرانسیل و انتگرال در سطح دانشگاهی و مهندسی علاقه بیشتری نشان دادند.
خوشه ۲	رکورد شماره ۱۳	به نظر می‌رسد کاربران این خوشه جزء افراد تازه وارد سایت هستند که در حال جستجوی صفحات ابتدایی سایت بوده و یا جز کاربران پیشین هستند که تازه به سایت وارد شده و قصد مرور آن را دارند و وقت زیادی را در صفحات ابتدایی صرف می‌کنند.
خوشه ۳	رکورد شماره ٤٦	کاربرانی که قصد دسترسی به کتاب‌های علمی و دانشگاهی را دارند، بیشتر در این خوشه متمرکز هستند.
خوشه ٤	رکورد شماره ١٦٦	کاربرانی که در قسمت پروفایل خود مشغول اعمال تغییرات هستند و یا قصد سفارش و پیگیری کتاب مورد نظر خود را دارند.
خوشه ٥	رکورد شماره ١٨٦	ظاهراً کاربران این خوشه تمایل به همکاری و نشر کارهای خود از طریق این انتشارات را دارند.
خوشه ٦	رکورد شماره ۲۰۸	کاربران این خوشه بیشتر به دنبال کتاب‌های ریاضی مربوط به دسته آموزش و پرورش (به خصوص ابتدایی) بوده‌اند.
خوشه ۷	رکورد شماره ۲٤۲	کاربران این خوشه علاقه‌مند به کتاب‌های مختلف با موضوع ریاضی بوده‌اند.
خوشه ۸	رکورد شماره ۳۰۲	در این خوشه کاربران به دنبال کتاب‌های المپیادی ریاضی بوده‌اند.
خوشه ۹	رکورد شماره ۳۰٤	کاربرانی که بیشتر تمایل به مطالب زیست‌شناسی (به خصوص کتاب‌های مربوط به المپیاد زیست‌شناسی که در دسته کتب آموزش و پرورش قرار می‌گیرد) داشته‌اند، در این خوشه قرار گرفته‌اند.
خوشه ۱۰	رکورد شماره ۳۱۷	به نظر می‌رسد کاربران این خوشه بیشتر تمایل به کسب اطلاعات کلی درباره کتاب‌ها و خرید حضوری از طریق نمایندگی‌های انتشارات داشته باشند.
خوشه ۱۱	رکورد شماره ٤۲۰	کاربرانی که قصد دسترسی به کتاب‌های آموزش و پرورش و دانشگاهی را دارند، بیشتر در این خوشه متمرکز هستند.

ادامه **جدول ۲**

شماره خوشه	مرکز خوشه	نتایج حاصل از تحلیل
خوشه ۱۲	رکورد شماره ٤۲۱	این کاربران بیشتر به دنبال کتب مربوط به مقطع آموزش و پرورش بوده‌اند.
خوشه ۱۳	رکورد شماره ٥٥٤	کاربران این خوشه بیشتر تمایل به کسب اطلاعات کلی درباره کتاب‌ها و مرور کاتالوگ کتاب‌ها دارند و سپس به صفحه راهنمای خرید کتاب مراجعه کرده‌اند که این نشان می‌دهد این کاربران پتانسیل خرید کتاب را دارند و به دنبال خرید هستند.
خوشه ۱٤	رکورد شماره ٥٦٤	کاربران این خوشه بیشتر به مطالب مرتبط با فیزیک علاقه‌مند هستند.

۴-۵- نتایج حاصل از مرحله توصیه

پس از مشخص شدن پروفایل هر خوشه، نوبت ارائه توصیه به داده‌های مجموعه آزمایشی است. همان‌طور که در جدول ۱ مشخص است، ۲۵۵ داده آزمایشی وجود دارد که توسط الگوریتم شناسایی نزدیک‌ترین همسایه‌ها (که در آن k، ٥ فرض شده)، ٥ خوشه نزدیک به تک‌تک این داده‌های آزمایشی مشخص می‌شوند و پروفایل این خوشه‌های نزدیک به داده مورد نظر با هم تلفیق شده و به عنوان صفحات پیشنهادی به کاربر یا داده آزمایشی ارائه می‌شوند. در این پژوهش ٤ صفحه برای پیشنهاد به کاربر در نظر گرفته شده است. پروفایل خوشه‌های نزدیک‌تر به داده مورد نظر با هم تجمیع شده و این کار تا زمانی ادامه می‌یابد که چهار صفحه منحصر به فرد به عنوان صفحات پیشنهادی برای داده مورد نظر انتخاب شوند. در نهایت برای هر داده آزمایشی ٤ صفحه منحصر به فرد پیشنهاد داده می‌شود.

۵- ارزیابی کارایی مدل

پس از مشخص شدن پروفایل و صفحات پیشنهادی برای هر داده آزمایشی می‌توان مدل ارائه شده را به کمک معیارهایی مانند دقت و فراخوانی و معیار F[٦٤] ارزیابی کرد. در جدول ۳ تمامی این معیارها برای مدل ارائه شده در این پژوهش محاسبه شده است. این مقادیر نشان می‌دهند که حتی با وجود اعمال محدودیت در تعداد صفحات پیشنهادی به کاربر (٤ صفحه

پیشنهادی) و نیز محدودیت عدم استفاده از صفحه اصلی سایت به عنوان صفحه پیشنهادی، مدل ارائه شده قادر به پیش‌بینی برخی صفحات مورد استفاده کاربران جدید است. در نتیجه به کاربران برای داشتن پیمایش‌های لذت‌بخش‌تر و بهتر در سایت کمک می‌کند. در ادامه به منظور ارزیابی کارایی الگوریتم خوشه‌بندی مورد استفاده در مدل پژوهش (الگوریتم انتشار کشش) این مدل با مدلی مشابه که در قسمت خوشه‌بندی آن از الگوریتم پرکاربرد و محبوب کی- مینز استفاده شده مقایسه می‌شود.

نتایج یادشده در جدول ۳ حاکی از این است که مدل ارائه شده توسط الگوریتم انتشار کشش، علاوه‌بر این‌که نیازی به تعیین تعداد خوشه ندارد، عملکرد قابل قبولی دارد و به نسبت بهتر از الگوریتم پرکاربرد کی- مینز عمل می‌کند. دلیل پایین بودن کلی اعداد به‌دست‌آمده ایجاد محدودیت برای صفحات پیشنهادی، حذف صفحات با فرکانس کمتر از یک مقدار خاص و همچنین تنوع و گوناگونی داده‌هاست.

جدول ۳. ارزیابی عملکرد الگوریتم‌ها

الگوریتم	دقت	فراخوانی	معیار F
مدل ارائه شده در این پژوهش (با بهره‌گیری از الگوریتم انتشار کشش)	۰/۳۱۸۶	۰/۲۲۸۲	۰/۲۶۶۰
استفاده از الگوریتم کی- مینز (k=14) در گام خوشه‌بندی مدل ارائه شده در این پژوهش	۰/۲۸۹۲	۰/۲۰۷۲	۰/۲۴۱۴
استفاده از الگوریتم کی- مینز (k=5) در گام خوشه‌بندی مدل ارائه شده در این پژوهش	۰/۲۲۹۴	۰/۱۶۴۳	۰/۱۹۱۵
استفاده از الگوریتم کی- مینز (k=10) در گام خوشه‌بندی مدل ارائه شده در این پژوهش	۰/۲۶۵۷	۰/۱۹۰۳	۰/۲۲۱۸
استفاده از الگوریتم کی- مینز (k=20) در گام خوشه‌بندی مدل ارائه شده در این پژوهش	۰/۳۰۴۹	۰/۲۱۸۴	۰/۲۵۴۵

۶- نتیجه‌گیری

در این پژوهش مدلی جهت تحلیل رفتار کاربران الکترونیکی ارائه شده است. در این مدل نخست داده‌های کاربران گردآوری و پیش‌پردازش می‌شود. پس از شناسایی کاربران

منحصر به فرد و نشست‌های مربوط به آن‌ها، داده‌های مجموعه آموزشی به کمک الگوریتم خوشه‌بندی انتشار کشش که به تعداد خوشه‌ها وابسته نیست، خوشه‌بندی می‌شوند. در مرحله بعد در هر خوشه الگوریتم سی.ام. اسپید به منظور تحلیل رفتار کاربران آن خوشه و نیز یافتن الگوهای پرتکرار اعمال می‌شود. به کمک نتایج حاصل از این مرحله پروفایل‌های کاربری مربوط به هر خوشه استخراج می‌شوند. حال به منظور ارائه پیشنهاد صفحات به کاربران جدید به کمک الگوریتم همسایگان نزدیک، ۵ همسایه و خوشه نزدیک به کاربر جدید (مجموعه آموزشی) شناسایی می‌شود و پروفایل‌های این خوشه‌های نزدیک با هم تلفیق شده و تا سقف چهار صفحه به کاربر جدید توصیه می‌شود. در نهایت به کمک این صفحات پیشنهادی و نیز دنباله واقعی پیمایش کاربران می‌توان معیارهایی از قبیل دقت و فراخوانی و معیار F را محاسبه و کارایی مدل ارائه شده را ارزیابی کرد. نتایج قسمت ارزیابی نشان‌دهنده این است که مدل پیشنهادی این پژوهش نسبت به مدل پیاده‌سازی شده توسط الگوریتم کی- مینز کارایی قابل قبولی دارد و در صورت کاهش محدودیت‌ها و نیز به‌کارگیری داده‌های مربوط به بازه زمانی طولانی‌تر این کارایی بهبود می‌یابد.

۷- پی‌نوشت‌ها

1. Affinity Propagation
2. Co-occurrence Map Sequential Patterns Discovery using Equivalence classes(CM-SPADE)
3. World Wide Web
4. Web Content Mining
5. Web Structure Mining
6. Link
7. Structural Summery
8. Log File
9. Session
10. Data Preparation and Preprocessing
11. Pattern Discovery
12. Knowledge Discovery
13. Pattern Analysis
14. Knowledge Analysis and Presentation
15. Time-out
16. Density-based spatial clustering of applications with noise (DBSCAN)
17. Vector Analysis

18. Fuzzy Set Theory
19. CFWS:Clustering based on Frequent Word Sequences
20. K-means
21. Apriori
22. Sequence Mining
23. Footstep Graph
24. Back-propagation Network
25. User Navigation Profiles
26. Link Prediction
27. PAM:Partitioning around Medoid
28. SPADE:Sequential Patterns Discovery using Equivalence classes
29. Semantic
30. Stanford Topic Modelling Toolbox
31. Online
32. Self-Care
33. ART2-enhanced k-means
34. Sequence-based representation Schemes
35. Co-occurrence Map Sequential Patterns Discovery using Equivalence classes(CM-SPADE)
36. KNN: K-Nearest Neighbors
37. Data Cleaning and Filtering
38. De-Spidering
39. Bots
40. User Agent
41. URI:Universal Resource Identifier
42. IP
43. Reactive
44. Message-Passing
45. Representative
46. Real-Value
47. Similarity
48. Preferences
49. Self-Similarity
50. Responsibility
51. Availability
52. Similarity
53. Co-occurrence Pruning
54. Vertical Database Format
55. Performance
56. Horizontal Database Format
57. Sequences
58. Minimum Support

59. Itemset
60. Items
61. Association Rule Mining
62. Open Source
63. Minimum Support
64. F-Measure

۸- مراجع

[1] Varnagar, C. R., Madhak, N. N., Kodinariya, T. M., Rathod, J. N.,"Web Usage Mining: a Review on Process,Methods and Techniques", Information Communication and Embedded Systems (ICICES), International Conference on. IEEE, 2013, pp.(40-46)

[2] Ghaderian, M., "Improving website user model automatically using semantics with domain specific concepts ", M.S. thesis, Faculty of computer engineering and information technology, Amirkabir university of technology, 2009

[3] Khosravai,M.,"Extracting knowledge from web using data mining techniques", M.S. Thesis, Faculty of Industrial Engineering, K.N. Toosi university of technology, 2011

[4] Brufar, A., Rezaeyan, A., Shokuhyar, S., "Identifying the customer behavior model in life insurance Sector using data mining", Management Researcheses in Iran, 20(4), 2016, pp. 64-94

[5] Mazaheri HosseinAbadi, E., Talebpour, A., Rezaeyan, A., "Identification Power Nodes in Social Networks Using Data Mining", Management Researcheses in Iran, 19(2), 2015, pp. 161-182

[6] Guerbas, A., Addam, O., Zaarour, O., Nagi, M., Elhajj, A., Ridley, M., Alhajj, R.,"Effective Web Log Mining and Online Navigational Pattern Prediction", knowledge-based systems, 49, 2013, pp.50-62

[7] Song , Q. , Shepperd, M.,"Mining Web Browsing patterns for e-commerce", Computers in industry, 57, 2006, pp. 622-630

[8] Taherizadeh, S., Moghadam, N.,"Integrating Web Content Mining into Web Usage Mining for Finding Patterns and Predicting Users Behavior",International Journal of Information Science and management, 7, 2009, pp.51-66

[9] Chou,P.,li,P.,Chen, K.,Wu,M.,"Integrating Web Mining and Neural Network for Personalized e-commerce Automatic Service",Expert Systems with applications,37, 2010, pp.2898-2910

[10] Carmona, C.J. , Gallego, S. , Torres, F. , Bernal, E., Jesus, M.J. , Garcia, S. ,"Web Usage Mining to Improve the Design of an e-commerce Website: OrOliveSur.com", Expert Systems with Applications, 39 , 2012, pp. 11243-11249

[11] Arbelaitz, O., Gurrutxaga, I., Lojo, A., Muguerza, J. , MariaPerez, J., Perona, I.,"Web Usage and Content Mining to Extract Knowledge for Modelling the Users of the Bidasoa Turismo Web Site and to Adapt It", Expert Systems with applications, 40, 2013, pp.7478 -7491

[12] Kim, J.K. , Kim, S.H. ,"A personalized recommender system based on web usage mining and decision tree induction", Expert Systems with applications,23(3), 2002, pp.329- 342

[13] Park, S. , Suresh, C. , Jeong, B.K.,"Sequence-based clustering for Web usage mining A new experimental framework and ANN-enhanced k-means algorithm", Data and knowledge engineering, 65(3), 2008, pp.512-543

[14] Loyola, P., Roman, P.E., Velasquez, J.D. ,"Predicting web user behavior using learning-based ant colony optimization", Engineering

applications of artificial intelligence,25(5), 2012, pp.889-897

[15] PYadav, M., Feeroz, M., KYadav, V.,"Mining the customer behavior using web usage mining in e-commerce", Computing communications and networking technologies(ICCCNT), International conference on IEEE, 2012, pp.1-5

[16] Hung, Y.S., Chen, K.B , Yang, C.T. , Deng, G.F., "Web usage mining for analyzing elder self-care behavior patterns", Expert Systems with applications, 40(2),2013,pp.775-783

[17] Ghelichkhani, B.," Combination of web usage and web content mining for personalization based on recommendation", M.S. thesis, Faculty of engineering, Tarbiat ModarresUniversity, 2011

[18] Lopes, P., Roy, B.,"Dynamic Recommendation System Using Web Usage Mining for e-commerce Users", Procedia Computer Science, 45, 2015, pp.60-69

[19] Sohrabi, B., Raesi Vanani, E., Zare Mirakababd, F., "Designing a Recommender System for Optimizing and Managing Bank Facilities through the Utilization of Clustering and Classification Algorithms", Modern Researches in Decision Making,1(2), 2016, pp.53-76

[20] Frey, B.J, Duek, D.,"Clustering by passing messages between data points", Science, 315(5814), 2007, 972-976

[21] Viger, P.F , Gomariz, A., Campos, M., Thomas, R.," Fast vertical mining of sequential patterns using co-occurrence information", Pacific-asia conference on knowledge discovery and data mining, Springer International publishing, 2014, pp.40-52

[22](Online)Available at http://www.philippe-fournier viger.com/spmf/index.php, 2017