مجله تحقیق در عملیات در کاربردهای آن، سال یازدهم، شماره دوم (پیاپی ٤١)، تابستان ٩٣، ص ص ١٥٢-١٣٩
شاپا ٧٢٨٦-٢٢٥١

روشی برای بهبود عملکرد سیستم‌های پیشنهادگر توریسم

اعظم آقایی*[1]، رضا رافع[2]، مهرگان مهدوی[3]

۱- دانشجوی کارشناسی ارشد، دانشگاه آزاد اسلامی واحد ملایر، گروه مهندسی کامپیوتر، ملایر، ایران

۲- استادیار، دانشگاه اراک، گروه مهندسی کامپیوتر، اراک، ایران

۳- استادیار، دانشگاه گیلان، گروه مهندسی کامپیوتر، ایران

رسید مقاله: ۱۷ آذر ۱۳۹۲

پذیرش مقاله: ۲۲ اردیبهشت ۱۳۹۳

چکیده

سیستم‌های پیشنهادگر سیستم‌های هوشمندی هستند که با تحلیل رفتار کاربران با شیوه‌های مختلف مانند داده‌کاوی، اقدام به پیشنهاد مناسب‌ترین کالا برای آنان می‌نمایند. این سیستم‌ها رویکردی هستند که برای مواجهه با مشکلات ناشی از حجم فراوان و رو به رشد اطلاعات ارایه شده‌اند و به یک کاربر کمک می‌کنند تا در میان حجم عظیم اطلاعات، سریع‌تر به هدف خود یعنی رسیدن به گزینه مفید و مورد علاقه، نزدیک شود در این مقاله مدلی برای افزایش کیفیت پیشنهاددهی در سیستم‌های پیشنهادگر توریسم ارایه می‌گردد. در این مدل با ترکیب ماتریس رتبه‌دهی کاربران و ماتریس حاصل از اطلاعات شخصی کاربران به یک تابع شباهت جدید دست می‌یابیم که محدوده همسایگی بهتری را برای کاربران مشخص می‌کند و در نتیجه باعث بالا رفتن کیفیت پیشنهاد می‌شود و از طرفی چون تابع شباهت جدید فقط وابسته به ماتریس نرخ گذاری نیست، در مواردی که کاربر به آیتمی نرخ نداده باشد می‌توان شباهت را از طریق ماتریس مشخصات کاربر به‌دست آورد و مانع از بروز مشکل شروع سرد که یکی از چالش‌های موجود در سیستم‌های پیشنهادگر است شد.

کلمات کلیدی: سیستم‌های پیشنهادگر، فیلترینگ مشارکتی، شباهت، داده کاوی، نرخ گذاری.

۱ مقدمه

حجم فراوان و رو به رشد اطلاعات بر روی اینترنت، فرایند تصمیم گیری و انتخاب اطلاعات و یا کالاهای مورد نیاز را برای بسیاری از کاربران اینترنت دشوار کرده است. این موضوع انگیزه‌ای شد تا محققین را وادار به پیداکردن راه حلی برای رویارویی با این مشکل اساسی عصر جدید که با عنوان سربار اطلاعاتی شناخته می‌شود، نماید. سربار اطلاعاتی به حالتی گفته می‌شود که میزان اطلاعات برای تصمیم‌گیری درمورد یک موضوع بیش از حد زیاد است [۱].

* عهده‌دار مکاتبات

آدرس الکترونیکی: aghaee_a@liau.ac.ir

تا به امروز روش‌های گوناگونی برای حل مشکل سربار اطلاعات پیشنهاد شده است. یکی از این رویکردها، استفاده از موتورهای جستجو است؛ اما تاکنون این موتورها در شخصی سازی نتایج جستجو چندان موفق نبوده‌اند و اغلب نتایج یکسانی را برای همه کاربران برمی گردانند در حالی که ممکن است دو کاربر، پروفایل‌های کاملاً متفاوتی داشته باشند و جنبه‌های مختلفی از نتایج جستجو را مد نظر قرار داده باشند [۲].

روش کاربردی دیگر، استفاده از سیستم‌های پیشنهادگر است. سیستم‌های پیشنهادگر ابزاری هستند که برای کمک کردن به این فرایند اجتماعی طبیعی و مقابله با این" سربار اطلاعات"طراحی شده‌اند. وظیفه آن‌ها این است که از میان حجم انبوهی از آیتم هائی که هر روزه ایجاد می‌شوند، تنها تعداد اندکی را جدا کنند که ممکن است مورد علاقه کاربر خاصی قرار بگیرد و توجه وی را به خود جلب کند. جای تعجب نیست که سیستم‌هایی که انجام این فرایند را خودکار می‌کنند، در فضای اینترنت محبوبیت زیادی پیدا کرده‌اند [۳].

در سیستم‌های پیشنهادگر تلاش براین است تا با حدس زدن شیوه تفکر کاربر (به کمک اطلاعاتی که از نحوه رفتار وی با کاربران مشابه وی و نظرات آن‌ها) به وی مناسب‌ترین و نزدیک‌ترین کالا به سلیقه او را شناسایی و پیشنهاد کنیم. این سیستم‌ها در حقیقت همان فرایندی هستند که ما در زندگی روزمره خود در طی آن تلاش می‌کنیم تا افرادی با سلایق نزدیک به خود را پیدا کرده و از آن‌ها در مورد انتخاب‌هایمان نظر بخواهیم [۴و۵].

۲ روش‌های پیشنهاددهی

سامانه‌های پیشنهادگر، سامانه‌های پردازش اطلاعات هستندکه انواع گوناگون داده را برای ایجاد پیشنهاداتشان جمع‌آوری می‌کنند. این سامانه‌ها با توجه به روش‌هایی که برای پالایش اطلاعات استفاده می‌کنند، به دو روش اصلی زیر تقسیم بندی می‌شوند[۶و۷].

۲-۱ روش مبتنی بر محتوا

ریشه‌های این روش به شاخه بازیابی و فیلترکردن اطلاعات باز می‌گردد. به سبب اهمیت و پیشرفت‌های زودهنگام در زمینه بازیابی اطلاعات و نیز به علت اهمیت کاربردهای آن در زمینه‌هایی که صرفا حاوی متن هستند، بسیاری از سیستم‌های مبتنی بر محتوای فعلی، بر آیتم هائی که دارای اطلاعات متنی هستند نظیر اسناد، وب سایت ها و پیام‌های خبری تمرکز دارند [۷]. در این رویکرد، ابتدا آیتم‌هایی که کاربر به آن‌ها امتیاز قابل قبولی داده است به دست می‌آیند. سپس در لیست کل آیتم‌ها به دنبال نمونه‌هایی که مشابه با نمونه‌های امتیازداده شده توسط کاربر هستند، می‌گردیم و از بین آن‌ها، شبیه‌ترین آیتم‌ها را به کاربر پیشنهاد می‌کنیم. بنابراین این متد به دو نوع از اطلاعات نیازمند است؛ یکی اطلاعاتی درباره پروفایل کاربر و دیگری اطلاعاتی درباره محتوای آیتم‌ها.

۲-۲ روش فیلترینگ مشارکتی

در این سیستم‌ها، به جای استفاده از محتوای آیتم‌ها، از نظرات یا نرخ گذاری‌های انجام شده توسط سایر کاربران برای ارایه پیشنهاد به کاربرهدف، استفاده می‌شود. در این رویکرد، لیست آیتم‌های پیشنهادی براساس این اصل که کاربرانی مشابه کاربر هدف از آن‌ها رضایت داشته‌اند تهیه می‌شود لذا تمرکز روی یافتن شباهت بین کاربران است[۸].

به عبارت دیگر، الگوریتم‌های این رویکرد بر این پایه استوارند کسانی که در گذشته با هم توافق داشته‌اند به احتمال زیاد در آینده نیز توافق خواهند داشت. یعنی کاربرانی که در گذشته رفتار مشابهی از خود ابراز داشته‌اند می‌توانند در مورد آیتم نرخ‌گذاری نشده نیز در مورد یکدیگر به عنوان پیشنهادگر رفتار کنند. عملکرد این سیستم به این صورت است که با تجزیه و تحلیل آماری اطلاعات و یا استخراج داده‌های کاربر، رفتار گذشته وی و سایر اطلاعات، یک محدوده همسایگی از افراد با سلایق و علایق مشترک ایجاد نموده و سپس با یافتن نزدیک‌ترین همسایه‌ها برای هر کاربر به پیشنهاد نرخ گذاری‌ها و انتخاب‌های این همسایه‌ها به کاربر هدف می‌پردازد. علاوه بر این فیلترینگ مشارکتی محدوده پیشنهادها را به موارد مشابهی که کاربر قبلا آن‌ها را نرخ گذاری کرده، محدود نمی‌کند و به عنوان جدیدترین و گسترده‌ترین تکنیک پیشنهاد مورد استفاده قرار می‌گیرد [۹و۱۰] و ما در این مقاله از همین تکنیک استفاده می‌کنیم.

۲-۲-۱ متدلوژی

یک سیستم بر مبنای فیلترکننده مشارکتی، یک ماتریس کاربر-آیتم دریافت می‌کند که سطرهای آن کاربران و ستون‌های آن آیتم‌ها هستند. هر درایه‌ی آن میزان علاقه‌ای است که یک کاربر نسبت به یک آیتم نشان داده است. این علاقه‌مندی می‌تواند به طور مستقیم از خود کاربر گرفته شود(مثلاً بین ۱ تا ۵ ستاره) و یا از روی آیتم‌های خریده شده یا سایر فعالیت‌های کاربر استدلال شود [۷]. درحقیقت می‌توان گفت این ماتریس کاربر-آیتم، داده‌ی ورودی کلیدی برای سیستم‌های مشارکتی محسوب می‌شود. نمونه ای از این ماتریس را در زیر مشاهده می‌نمایید که خانه‌های خالی بیانگر این است که کاربر به آیتم نرخی نداده است و می‌خواهیم عمل پیشگویی را در مورد آیتم ۶ برای کاربر ۳ انجام دهیم.

جدول ۱. یک نمونه از ماتریس کاربر-آیتم

	آیتم ۱	آیتم ۲	آیتم ۳	آیتم ۴	آیتم ۵	آیتم ۶
کاربر۱	۱		۴	۳	۱	
کاربر۲		۱	۳	۱	۲	۵
کاربر۳	۵	۲		۴		؟

هدف نهایی آن است که با گرفتن یک کاربر به عنوان ورودی، آیتمی را به وی پیشنهاد کنیم که احتمالاً مورد علاقه وی قرار می‌گیرد. الگوریتم زیر این هدف را برآورده می‌کند:

۱. محاسبه شباهت: ابتدا میزان شباهت همه کاربران با کاربر فعال(کاربری که قصد داریم برای وی پیش بینی انجام دهیم) مشخص می‌شود. برای این کار تکنیک های مختلفی پیشنهاد شده است. از جمله ضریب همبستگی پیرسون، معیار شباهت کسینوسی و ضریب همبستگی کسینوس اصلاح شده. که طبق مشاهدات صورت گرفته در اکثریت قریب به اتفاق پژوهش‌ها، بهترین آن‌ها ضریب همبستگی پیرسون بوده است. این ضریب از رابطه زیر محاسبه می‌شود:

$$\mathrm{sim}_{\mathrm{i,j}} = \frac{\sum_{m\in(\mathrm{i}\cap\mathrm{j})}(r_{\mathrm{i,m}} - \bar{r}_i)(r_{\mathrm{j,m}} - \bar{r}_j)}{\sqrt{\sum_{m\in(\mathrm{i}\cap\mathrm{j})}(r_{\mathrm{i,m}} - \bar{r}_i)^2}\sqrt{\sum_{m\in(\mathrm{i}\cap\mathrm{j})}(r_{\mathrm{j,m}} - \bar{r}_j)^2}} \qquad (۱)$$

در این رابطه، $sim_{i,j}$ شباهت دو کاربر i و j را بیان می‌کند. اگر i و j دو کاربر باشند، آن گاه $\mathrm{i}\cap\mathrm{j}$ مجموعه همه آیتم‌هایی است که هر دو کاربر i و j به آن رتبه داده‌اند. $\bar{r}_i$ و $\bar{r}_j$ میانگین رتبه‌های کاربر i و کاربر j روی همان آیتم‌های $\mathrm{i}\cap\mathrm{j}$ است. $r_{i,m}$ رتبه کاربر i به آیتم m و $r_{j,m}$ رتبه کابر j به آیتم m است.

۲. تخمین پیشنهاد: برای ایجاد پیشنهاد برای کاربر u روی آیتم i از فرمول زیر استفاده می‌شود[۱۱و۱۲]:

$$\mathrm{prediction}_{\mathrm{u,i}} = \frac{\sum_{n\in\mathrm{Neighbors}}\left(r_{\mathrm{n,i}} - \bar{r}_n\right)\mathrm{sim}_{\mathrm{u,n}}}{\sum_{n\in\mathrm{Neighbors}}\left|\mathrm{sim}_{\mathrm{u,n}}\right|} + \bar{r}_u \qquad (۲)$$

در رابطه بالا، Neighbors مجموعه نزدیک‌ترین کاربران به کاربر u است که انتخاب تعداد اعضای مجموعه Neighbors بستگی به اهمیت زمان اجرا و میزان کیفیت دارد. لیست شباهت‌های به دست آمده میان کاربر u و بقیه کاربران مرتب می‌شود و مجموعه ای از کاربران با بیش‌ترین ترین شباهت‌ها به عنوان Neighbors انتخاب می‌شوند. $\bar{r}_u$ و $\bar{r}_n$ میانگین رتبه‌ها برای کاربر u و کاربر n روی همه آیتم‌ها به جز آیتم i هستند و $\mathrm{sim}_{\mathrm{u,n}}$ همان شباهت میان دو کاربر u و n است.

۳ سیستم‌های پیشنهادگر در گردشگری الکترونیکی

یکی از مهم‌ترین عوامل موفقیت یک سیستم پیشنهادگر گردشگری الکترونیکی و افزایش رضایتمندی گردشگران، هوشمندی این سیستم است [۱۳] یک سیستم هوشمند می‌تواند ساختار و محتوای خود را بر مبنای تحلیل رفتار کاربر بهبود دهد یعنی سیستم پیشنهادگر گردشگری الکترونیکی باید تا حد زیادی کاربر پسند بوده، به آسانی قابل مدیریت باشد[۱۴]. به عنوان مثال، گردشگران بر مبنای پارامترهایی چون سن، جنس، علایق و ترجیحات و ... شناسایی می‌شوند. این نوع هوشمندی می‌تواند با الگوریتم‌های هوش مصنوعی و روش‌های داده کاوی و وب کاوی شبیه سازی شود[۱۵].

با توجه به اهمیت فوق العاده ای که مقوله گردشگری از دیدگاه اقتصادی، فرهنگی، سیاسی و اجتماعی برای کشورها دارد، ضروری به نظر می‌رسد که کشورهای مختلف که دارای پتانسیل‌های گردشگری هستند در

قالب طرح‌های تحقیقاتی و کاربردی به طراحی و پیاده سازی ساختارهای توریسم الکترونیکی بپردازند و نتایج پژوهش‌ها را تبدیل به برنامه‌های کاربردی متناسب با موقعیت خود نمایند[۱۶]. حاصل نهایی این طرح‌ها منجر به توسعه سریع تر صنعت گردشگری با تکیه بر امکانات فناوری اطلاعات، افزایش سطح رضایتمندی مشتریان، موفقیت برگزارکنندگان تور از نظر کمی، کیفی و موفقیت در عرصه رقابت‌های بین المللی تجاری خواهد بود.

مطالعات اخیر نشان می‌دهد که حداقل در کشورهای پیشرفته، وب منبع اصلی اطلاعات برای افرادی است که قصد مسافرت دارند. در نتیجه این حوزه جلودار تکنولوژی اطلاعات است و هنوز به عنوان یک حوزه تحقیقاتی جذاب محسوب می‌شود که پتانسیل‌های استخراج نشده زیادی در آن وجود دارد و در این زمینه کاربردهای پیشنهادگر می‌توانند ابزارهای ارزشمندی باشند[۱۵].

با استفاده از فن آوری اطلاعات، ارایه سرویس های موردنیاز گردشگران ساده تر، با کیفیتی بالاتر و هزینه ای کمتر انجام می‌پذیرد. در واقع گردشگری الکترونیکی ارایه الکترونیکی کلیه خدماتی است که در گذشته، گردشگران به صورت سنتی از آن‌ها استفاده می‌کردند به علاوه سرویس هایی که به واسطه فن آوری اطلاعات میسر شده‌اند[۱۷].

دلیل اینکه چرا کاربرد موفق تکنیک های پیشنهاد فیلم یا کتاب نمی‌توانند مستقیما در حوزه جهانگردی بکار برده شوند این است که مدل کردن پروفایل های کاربری توریسم به طور دقیق، نسبت به حوزه‌های کاربردی دیگر کار دشواری است[۱۸]. گردشگری فعالیتی است که تکرار کمتری نسبت به کارهایی مانند فیلم یا خرید کتاب دارد. بنابراین تعداد آیتم‌های گردشگری نرخ گذاری شده در دسترس نسبت به سایر آیتم‌ها خیلی کمتر است. به بیان دیگر ساختار یک محصول گردشگری پیچیده تر از ساختار یک فیلم یا یک کتاب است. از طرفی تکنیک های فیلترینگ همکارگونه وقتی به خوبی کار می‌کنند که اجتماع عظیمی از کاربران وجود داشته باشد و هر کاربری تعداد معینی از آیتم‌ها را از قبل نرخ گذاری کرده باشد. طرح‌های مسافرت منحصر به فرد به دلیل اینکه تکرار کمتری دارند و به علاوه ممکن است آیتم‌ها ساختار پیچیده تری داشته باشند در نتیجه ایجاد پروفایل های معقول کاربری مشکل خواهد بود.[۱۹].

در پیشنهادات جهانگردی علاوه بر اینکه باید علایق کاربر و یا گروه کاربران در نظر گرفته شود باید اطلاعات محیطی شامل مسافت میان مکان‌ها، زمان مسافرت، تسهیلات حمل و نقل و ... نیز در نظر گرفته شود. با وجود همه این مشکلات، علاقه زیادی به استفاده از سرویس های پیشنهادگر مسافرتی وجود دارد و تیم های تحقیقاتی زیادی در این زمینه مشغول به فعالیت هستند[۲۰].

٤ روش پیشنهادی

در این بخش به معرفی مدلی برای افزایش کیفیت پیشنهاددهی در سیستم‌های پیشنهادگر توریسم پرداخته می‌شود. در این مدل با ترکیب ماتریس رتبه‌دهی کاربران و ماتریس حاصل از اطلاعات شخصی کاربران

به یک تابع شباهت جدید دست می‌یابیم که محدوده همسایگی بهتری را برای کاربران مشخص می کند و در نتیجه باعث بالا رفتن کیفیت پیشنهاد می‌شود.

٤-١ چگونگی تأثیرگذاری فاکتور های مختلف در سیستم‌های پیشنهادگر توریسم

تحلیل‌ها نشان می‌دهد که کیفیت پیشنهادات وابسته به تراکم ماتریس نرخ گذاری است. افزایش تراکم این ماتریس با استفاده از ترکیب همه منابع اطلاعاتی در دسترس، ممکن خواهد بود[۳].

٤-١-١ سازماندهی همسایه‌ها

در اولین مرحله با ارزیابی روابط ماتریس کاربر- آیتم باید زیر مجموعه ای از نزدیک‌ترین کاربران به کاربر هدف (شبیه‌ترین کاربران به کار هدف) را جستجو کرد. در این راستا، نمایه هر جفت کاربر b و a، به منظور ارزیابی میزان شباهت آن‌ها، با یکدیگر مقایسه می‌شوند اگر $sim_{a,b}$ بیانگر میزان شباهت a و b باشد مقدار آن معمولاً دربازه ۱ (شباهت کامل) تا ۱- (تفاوت کامل) تغییر می کند. اگر در نمایه از کاربران هیچ اشتراکی وجود نداشته باشد معیاری برای تشخیص میزان شباهت وجود ندارد در نتیجه مقدار شباهت برابر با صفر می‌شود. در این مرحله برای به‌دست آوردن یک تابع شباهت دقیق تر، ضریب تاثیرفاکتورهای مختلف در نرخ گذاری را به‌دست می‌آوریم و به تابع شباهت اولیه که از ماتریس کاربر-آیتم به‌دست آمده اضافه می کنیم.

برای اینکار از نرم افزار داده کاوی کلمنتاین استفاده کرده و با استفاده از الگوریتم‌های ساخت درخت موجود در این نرم افزار ضریب تاثیر فاکتورهای مختلف را به‌دست می‌آوریم.

٤-٢ مجموعه داده

مجموعه داده ما در این تحقیق شامل ۲۷۶ کاربر، ۴۱ شهر و ۳۸۳۶ رتبه کاربران می‌باشد که از طریق پرسش‌نامه جمع‌آوری شده است. در این مجموعه داده هر کاربر رتبه‌ای بین ۱ تا ۵ به شهرها داده است که رتبه بالاتر نشان‌دهنده علاقه‌مندی بیشتر است [۱۵].

جدول ۲. نمونه‌ای از مجموعه داده

user ID	coor X	coor Y	age	married	Gender	Occupation	Children	city name	rating
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	Student	۰	albacete	۲
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	۰	alicante	۲
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	۰	almeria	۳
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	۰	asturias	۲
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	۰	Avila	۲
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	۰	badajoz	۳
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	۰	baleares	۳
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	۰	barcelona	۳
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	۰	basque	۳
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	۰	burgos	۴

user ID	coor X	coor Y	age	married	Gender	Occupation	Children	city name	rating
٢	٧٢٩٣٩٤/٤١٦	١٤٥٩٧٧٨٦/٩٤	٢٨	yes	Male	scientist	٢	almeria	٢
٢	٧٢٩٣٩٤/٤١٦	١٤٥٩٧٧٨٦/٩٤	٢٨	yes	Male	scientist	٢	asturias	٣
٢	٧٢٩٣٩٤/٤١٦	١٤٥٩٧٧٨٦/٩٤	٢٨	yes	Male	scientist	٢	avila	٣
٢	٧٢٩٣٩٤/٤١٦	١٤٥٩٧٧٨٦/٩٤	٢٨	yes	Male	scientist	٢	badajoz	٢
٢	٧٢٩٣٩٤/٤١٦	١٤٥٩٧٧٨٦/٩٤	٢٨	yes	Male	scientist	٢	baleares	٢
٢	٧٢٩٣٩٤/٤١٦	١٤٥٩٧٧٨٦/٩٤	٢٨	yes	Male	scientist	٢	barcelona	٢
٢	٧٢٩٣٩٤/٤١٦	١٤٥٩٧٧٨٦/٩٤	٢٨	yes	Male	scientist	٢	basque	٢
٢	٧٢٩٣٩٤/٤١٦	١٤٥٩٧٧٨٦/٩٤	٢٨	yes	Male	scientist	٢	burgos	٤
٢	٧٢٩٣٩٤/٤١٦	١٤٥٩٧٧٨٦/٩٤	٢٨	yes	Male	scientist	٢	caceres	٤
٢	٧٢٩٣٩٤/٤١٦	١٤٥٩٧٧٨٦/٩٤	٢٨	yes	Male	scientist	٢	cadiz	٣
٢	٧٢٩٣٩٤/٤١٦	١٤٥٩٧٧٨٦/٩٤	٢٨	yes	Male	scientist	٢	cantabria	٢
٣	٧٢٩١٥٩/٩٥١	١٤٥٩٧٧٥٣/٣٥	٦٠	no	Female	professor	٠	avila	٢
٣	٧٢٩١٥٩/٩٥١	١٤٥٩٧٧٥٣/٣٥	٦٠	no	Female	professor	٠	badajoz	٣
٣	٧٢٩١٥٩/٩٥١	١٤٥٩٧٧٥٣/٣٥	٦٠	no	Female	professor	٠	baleares	٤
٣	٧٢٩١٥٩/٩٥١	١٤٥٩٧٧٥٣/٣٥	٦٠	no	Female	professor	٠	barcelona	٣
٣	٧٢٩١٥٩/٩٥١	١٤٥٩٧٧٥٣/٣٥	٦٠	no	Female	professor	٠	basque	٣
٣	٧٢٩١٥٩/٩٥١	١٤٥٩٧٧٥٣/٣٥	٦٠	no	Female	professor	٠	burgos	٤
٣	٧٢٩١٥٩/٩٥١	١٤٥٩٧٧٥٣/٣٥	٦٠	no	Female	professor	٠	caceres	٤
٣	٧٢٩١٥٩/٩٥١	١٤٥٩٧٧٥٣/٣٥	٦٠	no	Female	professor	٠	cadiz	٢
٣	٧٢٩١٥٩/٩٥١	١٤٥٩٧٧٥٣/٣٥	٦٠	no	Female	professor	٠	cantabria	٢
٣	٧٢٩١٥٩/٩٥١	١٤٥٩٧٧٥٣/٣٥	٦٠	no	Female	professor	٠	castello	٣
٣	٧٢٩١٥٩/٩٥١	١٤٥٩٧٧٥٣/٣٥	٦٠	no	Female	professor	٠	ciudad real	٣

٤-٣ الگوریتم‌های ساخت درخت تصمیم

این الگوریتم‌ها تمامی صفات مجموعه داده را بررسی می‌کنند تا به صفتی برسند که بهترین رده‌بندی و پیش‌بینی را با تقسیم داده به زیرگروه‌ها انجام دهد. این صفات به طور بازگشتی تکرار می‌شود تا باز هم زیر گروه‌ها به گروه‌های دیگری شکسته شوند. صفات هدف یا ورودی می‌توانند از نوع عددی و یا طبقه‌ای بر حسب الگوریتم مورد استفاده باشند. اگر یک بازه مورد استفاده قرار گیرد نتیجه کار یک درخت رگرسیون خواهد بود، اما اگر ورودی‌ها به صورت رده ای باشند نتیجه کار یک درخت رده بندی خواهد بود.

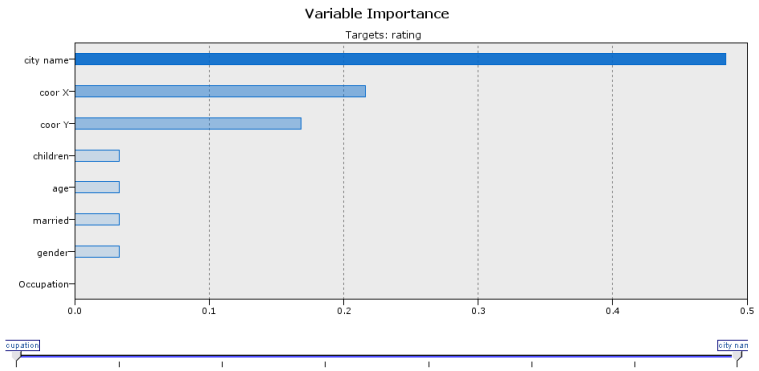
در مجموعه داده مورد استفاده در این تحقیق، چون فیلد هدف ما از نوع بازه ای است در نتیجه از بین ۴ مدل درختی موجود در نرم افزار کلمنتاین تنها ۲ مدل قابل استفاده بود، که عبارت‌اند از مدل C&R و مدل CHAID.

درخت تقسیم و رگرسیون(C&R) یک درخت تصمیم تولید می‌کند که سعی در پیش بینی رده بندی مشاهدات آینده دارد. این روش سعی در کم کردن ناخالصی در هر رده دارد. یک گره وقتی کاملاً عاری از ناخالصی است که تمامی عناصر یک زیرگروه آن متعلق به یک رده از فیلد هدف باشند. صفت پیش بینی‌کننده و فیلد هدف می‌توانند از دو نوع بازه‌ای و رده‌ای باشند. تمامی تقسیم‌بندی‌ها دودویی خواهد بود. به این معنی که فقط دو زیرگروه از هر گره منشعب خواهد شد.

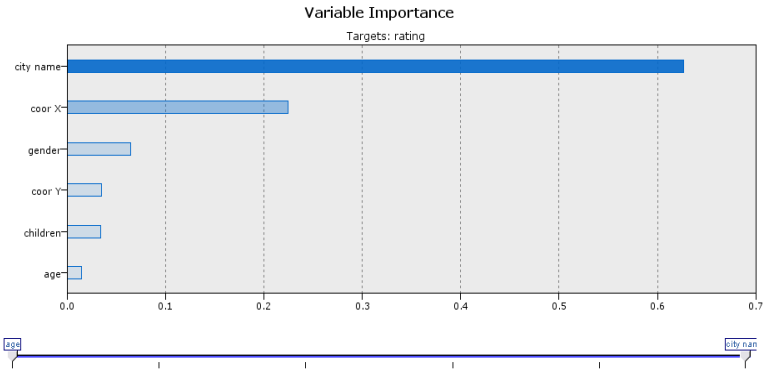
بر خلاف درخت C&R، درخت CHAID می‌تواند درختی تولید کند که در برخی موارد به صورت غیر دودویی عمل کند، یعنی یک گروه آن به سه زیرگروه و یا بیشتر شکسته شود. صفات پیش‌بینی کننده و هدف می‌توانند هم از نوع بازه‌ای و هم از نوع رده‌ای باشد.

دو مدل دیگر درخت تصمیم موجود در نرم افزار کلمنتاین را نتوانستیم مورد استفاده قرار دهیم چون فیلد هدف در آن‌ها باید حتما از نوع رده ای باشد در حالیکه در مجموعه داده ما فیلد هدف از نوع بازه ای است.

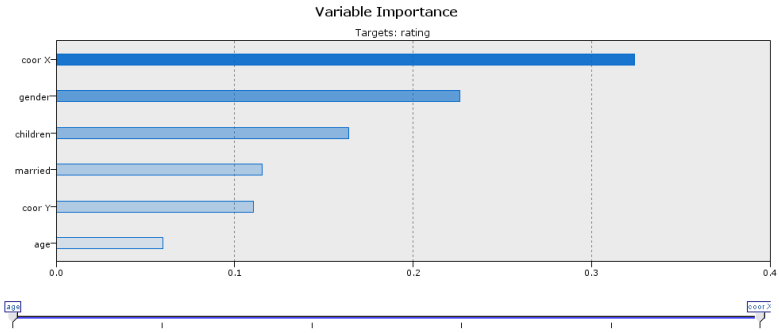
مجموعه داده مورد نظر توسط سه روش، مورد داده کاوی قرار گرفت که نتایج حاصل از آن‌ها در شکل‌های زیر قابل مشاهده است. در سه روش مورد آزمایش، مشاهده می‌شود که ضرایب تاثیر فاکتورها در مدل‌های مختلف، متفاوت است. مثلا ضریب تاثیر فاکتور موقعیت مکانی کاربر، با استفاده از مدل C&R برابر ۰/۴۸، مدل CHAID برابر ۰/۶۲ و شبکه عصبی برابر ۰/۳۲ به‌دست آمده است. حال برای اینکه ببینیم کدام مدل، ضرایب تاثیر را دقیق تر نشان داده باید از بخش آنالیز موجود در نرم افزار کلمنتاین کمک بگیریم .



شکل ۱. نتایج حاصل از داده کاوی با استفاده از درخت تصمیم *C&R*



شکل ۲. نتایج حاصل از داده کاوی با استفاده از درخت تصمیم *CHAID*



شکل ۳. نتایج حاصل از داده کاوی با استفاده از *Neural Network*

Results for output field rating

Individual Models

Comparing $R-rating with rating

'Partition'	1_Training	2_Testing
Minimum Error	-1.124	-0.7
Maximum Error	1.35	0.52
Mean Error	-0.036	-0.026
Mean Absolute Error	0.23	0.226
Standard Deviation	0.278	0.268
Linear Correlation	0.959	0.966
Occurrences	3,040	795

Comparing $R1-rating with rating

'Partition'	1_Training	2_Testing
Minimum Error	-2.262	-2.262
Maximum Error	2.6	2.286
Mean Error	-0.03	-0.022
Mean Absolute Error	0.58	0.582
Standard Deviation	0.709	0.717
Linear Correlation	0.647	0.655
Occurrences	3,040	795

Comparing $N-rating with rating

'Partition'	1_Training	2_Testing
Minimum Error	-2.187	-2.188
Maximum Error	2.01	2.01
Mean Error	-0.019	-0.0
Mean Absolute Error	0.65	0.675
Standard Deviation	0.914	0.935
Linear Correlation	0.039	-0.002
Occurrences	3,040	795

شکل ٤. نتایج حاصل از آنالیز سه مدل

نتایج حاصل از آنالیز مدل ها نشان می‌دهد که مدل درخت تصمیم C&R خطای کمتری را در پیشگویی‌ها دارد. بنابراین از ضرایب به‌دست آمده از این مدل برای رسیدن به یک تابع شباهت جدید استفاده کرده و برای نرمال سازی این ضرایب، رابطه زیر را به‌کار می‌بریم:

$$
\begin{aligned}
\text{ضریب} = {} & \left(\frac{\text{ضریب موقعیت مکانی}}{\left|\text{اختلاف موقعیت مکانی دو کاربر}\right|+1} \times \mathrm{C}\right) + \\
& \left(\frac{\text{ضریب سن}}{\left|\text{اختلاف سن دو کاربر}\right|+1} \times \mathrm{C}\right) + \\
& \left(\text{تعداد شهرها با نرخ یکسان} \times \text{ضریب شهر} \times C\right) + \\
& \begin{cases} \left(\text{ضریب ازدواج} \times C\right) & \text{:در صورت یکسان بودن وضعیت ازدواج} \\ \left(\text{ضریب ازدواج} \times 0\right) & \text{: در غیر اینصورت} \end{cases} + \\
& \begin{cases} \left(\text{ضریب جنسیت} \times C\right) & \text{: در صورت یکسان بودن جنسیت} \\ \left(\text{ضریب جنسیت} \times 0\right) & \text{: در غیر اینصورت} \end{cases} + \\
& \begin{cases} \left(\text{ضریب شغل} \times C\right) & \text{:در صورت یکسان بودن شغل} \\ \left(\text{ضریب شغل} \times 0\right) & \text{: در غیر اینصورت} \end{cases} +
\end{aligned}
\qquad (۳)
$$

C ضریب ثابتی است که جهت نرمال کردن رابطه بالا از آن استفاده می‌کنیم.

رابطه بالا را برای هر دو کاربر محاسبه کرده و ماتریس جدیدی به‌دست می‌آوریم. سپس این ماتریس را به ماتریس شباهت اولیه اضافه نموده در نتیجه ماتریس شباهت جدید به‌دست می‌آید و در رابطه (۲-۲)، ماتریس شباهت جدید را به جای ماتریس شباهت اولیه قرار می‌دهیم که باعث می‌شود مرز همسایگی‌ها تغییر کرده و کاربرانی که به کاربر هدف شبیه ترند به عنوان همسایه نزدیک‌تر انتخاب شوند، در نتیجه کیفیت پیشنهاد بالاتری خواهیم داشت و از طرفی چون تابع شباهت جدید فقط وابسته به ماتریس نرخ گذاری نیست، در مواردی که کاربر به آیتمی نرخ نداده باشد می‌توان شباهت را از طریق ماتریس مشخصات کاربر به‌دست آورد.

در این بررسی ۸۰ درصد داده‌ها به صورت تصادفی به عنوان مجموعه آموزش و ۲۰ درصد به عنوان مجموعه آزمایش در نظر گرفته شدند که نتایج حاصل از داده کاوی نشان داد که خطای حاصل از داده کاوی در مجموعه داده موردنظر، توسط درخت تصمیم C&R کمتر از دو روش دیگر است. بنابراین ضرایب به‌دست آمده از این مدل را مورد استفاده قرار دادیم.

در مرحله دوم روی همان مجموعه داده آزمایش، این بار از سیستم مشارکتی استفاده نموده و خطای حاصل را محاسبه می‌کنیم. برای اینکار ابتدا از روی ماتریس نرخ گذاری، شباهت میان کاربران را به‌دست می‌آوریم سپس با استفاده از تابع پیشگویی، رتبه شهرهایی که کاربر به آن‌ها نرخ نداده را تخمین می‌زنیم و در نهایت از طریق تابع ارزیابی میزان خطا را محاسبه می‌کنیم.

جدول ٣. بخشی از ماتریس حاصل از تابع شباهت اولیه

	user 1	user 2	user 3	user 4	user 5
user 1	١	٠/١۴٧۴	٠/٨۴٠٢	٠	−٠/١٢۵
user 2	٠/١۴٧۴	١	٠/٣٣٣	٠/٢٧۵٧	٠/٢٨٨٧
user 3	٠/٨۴٠٢	٠/٣٣٣	١	٠/۵٢٣٨	٠/١۶۶٧
user 4	٠	٠/٢٧۵٧	٠/۵٢٣٨	١	٠
user 5	−٠/١٢۵	٠/٢٨٨٧	٠/١۶۶٧	٠	١

در مرحله سوم از میزان اهمیت آیتم‌ها که در نرم افزار کلمنتاین به‌دست آمده بود برای تابع شباهت جدید به عنوان وزن استفاده کرده و دوباره خطا را برای همان مجموعه داده مورد محاسبه قرار می‌دهیم.

جدول ٤. بخشی از ماتریس حاصل از تابع شباهت جدید

	user 1	user 2	user 3	user 4	user 5
user 1	٢	١/٠٠۴٣	١/۴٩٠٣	٠/٧۶١۶	٠/۶٢۴٣
user 2	١/٠٠۴٣	٢	٠/٩٨۴١	١/٠٣٠٢	١/٠٣٨۶
user 3	١/۴٩٠١	٠/٩٨۴١	٢	١/٢٧۵٣	١/٠١٢۵
user 4	٠/۵٧١۶	١/٠٣٠٢	١/٢٧۵٣	٢	٠/۶۵٢۶
user 5	٠/۶٢۴٣	١/٠٣٨۶	١/٠١٢۵	٠/۶۵٢۶	٢

از مقایسه دو جدول مشاهده می‌شود که مرز همسایگی‌ها تغییر کرده و به علاوه در ماتریس شباهت جدید دیگر اثری از درایه صفر وجود ندارد.

در دو جدول زیر می‌توان تغییر ترتیب همسایگی‌ها را برای کاربران مشاهده نمود.

جدول ۵. ترتیب همسایگی در ماتریس شباهت اولیه

user 1	user 3	user 2	user 5	
user 2	user 3	user 4	user 5	user 1
user 3	user 1	user 4	user 2	user 5
user 4	user 3	user 2		
user 5	user 2	user 3	user 1	

جدول ٦. ترتیب همسایگی در ماتریس شباهت جدید

user 1	user 4	user 5	user 3	user 2
user 2	user 4	user 5	user 1	user 3
user 3	user 1	user 4	user 5	user 2
user 4	user 3	user 2	user 1	user 5
user 5	user 2	user 3	user 4	user 1

در این تحقیق به منظور پیاده سازی الگوریتم‌ها از برنامه MATLAB7.5 و برای اعمال روش‌های داده‌کاوی از نرم افزار SPSS Clementine12.0 استفاده شده است.

۵ ارزیابی مدل پیشنهادی

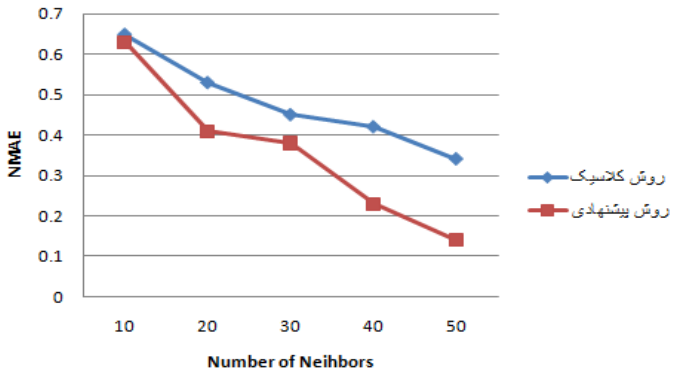
به منظور ارزیابی مدل پیشنهادی از معیار میانگین خطای مطلق نرمال شده (NMAE) استفاده می‌کنیم که پرکاربردترین معیار ارزیابی کیفیت در سیستم‌های پیشنهادگر می‌باشد. این معیار میانگین خطای میان پیشنهادات و رتبه‌های واقعی را برای یک مجموعه تست انتخاب شده محاسبه می‌کند[۳].

$$MAE = \frac{\sum_{(u,i)\in test} \left| prediction_{u,i} - real_{u,i} \right|}{n_{test}} \qquad (۴)$$

که در اینجا n_{test} تعداد اعضای مجموعه تست است. $prediction_{u,i}$ رتبه پیشگویی شده برای کاربر u و آیتم i است. $real_{u,i}$ رتبه واقعی کاربر u و آیتم i می‌باشد. NMAE به صورت زیر تعریف می‌شود[۳]:

$$NMAE = \frac{MAE}{r_{max} - r_{min}} \qquad (۵)$$

r_{max} و r_{min} بالاترین و پایین‌ترین محدوده رتبه در سیستم پیشنهادگر است.



شکل ۵. مقایسه روش پیشنهادی با روش کلاسیک

نمودار شکل ۵ رابطه بین NMAE با تعداد همسایه‌ها در روش پیشنهادی را با روش کلاسیک مقایسه می‌کند. همان‌طور که از نمودار مشخص است هرچه قدر تعداد همسایه‌ها بیشتر می‌شود NMAE در هر دو روش کاهش می‌یابد ولی در روش پیشنهادی، با به‌دست آوردن همسایه‌های بهتر برای کاربر هدف، توانستیم کیفیت پیشنهاد را بالاتر ببریم.

٦ نتیجه گیری

مهم ترین هدف سیستم‌های پیشنهادگر، پالایش حجم وسیعی از اطلاعات در جهت سازماندهی محتوای متناسب با علایق کاربر هدف است. در این تحقیق ما بر روی یکی از انواع سیستمهای پیشنهادگر به نام سیستم پیشنهادگر توریسم متمرکز شده و با ابداع یک روش جدید سعی در بهبود کارایی این نوع سیستم‌ها داشتیم.

روش پیشنهادی ما در دو فاز اندازه گیری شباهت و ارایه پیشنهادات، الگوریتم های کلاسیک پالایش مشارکتی را اصلاح کرده است. در راستای محاسبه شباهت، افراد از جنبه‌های مختلفی با یکدیگر مقایسه شده‌اند که از این بین، فاکتورهای اختلاف موقعیت مکانی دو کاربر، اختلاف سن دو کاربر و چند فاکتور دیگر مورد بررسی قرار گرفتند.

در ادامه به مقایسه روش پیشنهادی با روش‌های کلاسیک پالایش گروهی پرداختیم. در راستای این امر از معیار NMAE کمک گرفته شده است که بر روی مقایسه عددی نرخ های پیش بینی شده مجموعه داده با مقادیر واقعی وارد شده توسط کاربران متمرکز شده نتایج پیاده سازی روش پیشنهادی بر روی یک مجموعه داده نشان دهنده این امر است که از طریق تابع شباهت جدید می‌توان همسایه‌های بهتری را برای کاربر هدف به‌دست آورد و در نتیجه پیشگویی بهتری را انجام داد و از طرفی با مشکل شروع اولیه نیز مقابله نمود.

منابع

[1] Kai, Y., Schwaighofer, A., Tresp, V., Xu, X., (2011). Probabilistic Memory-Based Collaborative Filtering. IEEE Transactions on Knowledge and Data Engineering, 16: 56-69.

[2] Gong, S., Ye, H., (2009). An Item Based Collaborative Filtering Using BP Neural Networks Prediction. International Conference on Industrial and Information Systems, 146-148.

[3] Ricci, F., Rokach, L., Shapira, B., Recommender Systems Handbook 2010, Springer New York Dordrecht Heidelberg London.

[4] Neal, L., (2010). Evaluating Collaborative Filtering Over Time, Phd Thesis, Computer science. college london, london.

[5] Shanle, M., Li, X., Ding, Y., Orlowska, M., (2012). A Recommender System with Interest-Drifting. in Web Information Systems Engineering, France, 633-642.

[6] Alexandridis, G., Siolas, G., Stafylopatis, A., (2010). An Efficient Collaborative Recommender System Based on k-Separability. Computer Science, 6354: 198-207.

[7] Dong, S., Yen, D., Cheng Lin, H., (2011). Ming-Hung Shih, An implementation and evaluation of recommender systems for traveling abroad. Expert Systems with Applications, Science Direct, 38: 15344–15355.

[8] Beatriz, R., Molina, J., Pérez, F., (2012). Caballero, R., Interactive design of personalised tourism routes. Tourism Management, 33: p. 926e940.

[9] Linden, G., Smith, B., York, J., (2010). Amazon.com recommendations: item-to-item collaborative filtering, IEEE Internet Computing, 7: 76–80.

[10] Goy, A., Ardissono, L., Petrone, G., (2007). Personalization in E-Commerce Applications. Springer-Verlag, Berlin, Heidelberg, 485–520.

[11] Dickson, C., (2009). Towards ubiquitous tourist service coordination and process integration: A collaborative travel agent system architecture with semantic web services. Information Systems Frontiers, 241-256.

[12] Gretzel, U., (2011). Intelligent Systems In Tourism A Social Science Perspective. Annals of Tourism Research, 38: 757–779.

[13] Gülçin, B., Buse, E., (2011). Intelligent system applications in electronic tourism. Expert Systems with Applications, ScienceDirect, 38: 6586–6598.

[14] Helmut, B., Denk, M., Dittenbach, M., (2007). Photo-Based User Profiling for Tourism Recommender Systems. EC-Web, Springer-Verlag Berlin Heidelberg, 46–55.

[15] Inma, G., Sebastia, L., Onaindia, E., (2011). On the design of individual and group recommender systems for tourism. Expert Systems with Applications, Science Direct, 38: 7683–7692.
[16] Kabassi, K., (2010). Personalizing recommendations for tourists. Telematics and Informatics, 27: 51–66.
[17] Kawai, Y., Zhang, J., Kawasaki, H., (2009). Tour recommendation system based on web information and GIS. in Multimedia and Expo, IEEE International, 990-993.
[18] Laura, S., Girent, A., Garcia, I., (2010). A Multi Agent Architecture for Tourism Recommendation. Trends in PAAMS, AISC, Springer-Verlag Berlin Heidelberg, 71: p. 547–554.
[19] Prem, M., Mooney, R., Nagarajan, R., (2002). Content boosted collaborative filtering for improved recommendations. in 18th National Conference on Artificial Intelligence, Canada, Edmonton, 187–192.
[20] Montserrat, B., Moreno, A., Sánchez, D., Valls, A., (2012). Turist@: Agent-based personalised recommendation of tourist activities. Expert Systems with Applications, ScienceDirect, 39: 7319–7329.