

روشی برای بهبود عملکرد سیستم‌های پیشنهادگر تورسیم

اعظم آقایی^{۱*}، رضا رافع^۲، مهرگان مهدوی^۳

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 داده کاوی، 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 تورسیم 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 به دست 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 تصمیم‌گیری 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 پیدا کردن 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: aghae_a@liau.ac.ir

تا به امروز روش های 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 نبوده اند و اغلب نتایج یکسانی را برای همه کاربران برمی گردانند در حالی که ممکن است دو کاربر، پروفایل های کاملاً متفاوتی داشته باشند و جنبه های مختلفی از نتایج جستجو را مد نظر قرار داده باشند [۲].

روش کاربردی دیگر، استفاده از سیستم های پیشنهادگر است. سیستم های پیشنهادگر ابزاری هستند که برای کمک کردن به این فرایند اجتماعی طبیعی و مقابله با این "سربار اطلاعات" طراحی شده اند. وظیفه آنها این است که از میان حجم انبوهی از آیتم هایی که هر روزه ایجاد می شوند، 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 سیستم هایی که انجام این فرایند را خودکار می کنند، در فضای اینترنت محبوبیت زیادی پیدا کرده اند [۳].

در سیستم های 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 مناسب ترین و نزدیک ترین کالا به سلیقه او را شناسایی و پیشنهاد کنیم. این سیستم ها در حقیقت همان فرایندی هستند که ما در زندگی روزمره خود در طی آن تلاش می کنیم تا افرادی با سلیقه نزدیک به خود را پیدا کرده و از آنها در مورد انتخاب هایمان نظر بخواهیم [۴ و ۵].

۲ روش های پیشنهاددهی

سامانه های پیشنهادگر، سامانه های پردازش اطلاعات هستند که انواع گوناگون داده را برای ایجاد پیشنهادشان جمع آوری می کنند. این سامانه ها با توجه به روش هایی که برای پالایش اطلاعات استفاده می کنند، به دو روش اصلی زیر تقسیم بندی می شوند [۶ و ۷].

۲-۱ روش مبتنی بر محتوا

ریشه های این روش به شاخه بازیابی و فیلتر کردن اطلاعات باز می گردد. به سبب اهمیت و پیشرفت های زود هنگام در زمینه بازیابی اطلاعات و نیز به علت اهمیت کاربردهای آن در زمینه هایی که صرفاً حاوی متن هستند، بسیاری از سیستم های مبتنی بر محتوای فعلی، بر آیتم هایی که دارای اطلاعات متنی هستند نظیر اسناد، وب سایت ها و پیام های خبری تمرکز دارند [۷]. در این رویکرد، ابتدا آیتم هایی که کاربر به آنها امتیاز قابل قبولی داده است به دست می آیند. سپس در لیست کل آیتم ها به دنبال نمونه هایی که مشابه با نمونه های امتیاز داده شده توسط کاربر هستند، می گردیم و از بین آنها، شبیه ترین آیتم ها را به کاربر پیشنهاد می کنیم. بنابراین این متد به دو نوع از اطلاعات نیازمند است؛ یکی اطلاعاتی درباره پروفایل کاربر و دیگری اطلاعاتی درباره محتوای آیتم ها.

۲-۲ روش فیلترینگ مشارکتی

در این سیستم‌ها، به جای استفاده از محتوای آیت‌ها، از نظرات یا نرخ گذاری‌های انجام شده توسط سایر کاربران برای ارایه پیشنهاد به کاربر هدف، استفاده می‌شود. در این رویکرد، لیست آیت‌های 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 آیت‌ها نرخ گذاری 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 سلاقی و علائق 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 قبلاً آن‌ها را نرخ گذاری کرده، محدود نمی‌کند و به عنوان جدیدترین و گسترده‌ترین تکنیک پیشنهاد مورد استفاده قرار می‌گیرد [۱۰۹] و ما در این مقاله از همین تکنیک استفاده می‌کنیم.

۲-۲-۱ متدولوژی

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

جدول ۱. یک نمونه از ماتریس کاربر-آیت‌ها

	آیت ۱	آیت ۲	آیت ۳	آیت ۴	آیت ۵	آیت ۶
کاربر ۱	۱		۴	۳	۱	
کاربر ۲		۱	۳	۱	۲	۵
کاربر ۳	۵	۲		۴		۴

هدف نهایی آن است که با گرفتن یک کاربر به عنوان ورودی، آیت‌های 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 می شود. 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 پژوهش ها، بهترین آن ها ضریب همبستگی پیرسون بوده است. این ضریب از رابطه زیر محاسبه می شود:

$$\text{sim}_{i,j} = \frac{\sum_{m \in (i \cap j)} (r_{i,m} - \bar{r}_i)(r_{j,m} - \bar{r}_j)}{\sqrt{\sum_{m \in (i \cap j)} (r_{i,m} - \bar{r}_i)^2} \sqrt{\sum_{m \in (i \cap j)} (r_{j,m} - \bar{r}_j)^2}} \quad (1)$$

در این رابطه، $\text{sim}_{i,j}$ شباهت دو کاربر i و j را بیان می کند. اگر i و j دو کاربر باشند، آن گاه $i \cap j$ مجموعه همه آیتم هایی است که هر دو کاربر i و j به آن رتبه داده اند. $\bar{r}_i$ و $\bar{r}_j$ میانگین رتبه های کاربر i و کاربر j روی همان آیتم های $i \cap j$ است. $r_{i,m}$ رتبه کاربر i به آیتم m و $r_{j,m}$ رتبه کاربر j به آیتم m است. ۲. تخمین پیشنهاد: برای ایجاد پیشنهاد برای کاربر u روی آیتم i از فرمول زیر استفاده می شود [۱۱ و ۱۲]:

$$\text{prediction}_{u,i} = \frac{\sum_{n \in \text{Neighbors}} (r_{n,i} - \bar{r}_n) \text{sim}_{u,n}}{\sum_{n \in \text{Neighbors}} |\text{sim}_{u,n}|} + \bar{r}_u \quad (2)$$

در رابطه بالا، Neighbors مجموعه نزدیک ترین کاربران به کاربر u است که انتخاب تعداد اعضای مجموعه Neighbors بستگی به اهمیت زمان اجرا و میزان کیفیت دارد. لیست شباهت های به دست آمده میان کاربر u و بقیه کاربران مرتب می شود و مجموعه ای از کاربران با بیش ترین شباهت ها به عنوان Neighbors انتخاب می شوند. $\bar{r}_u$ و $\bar{r}_n$ میانگین رتبه ها برای کاربر u و کاربر n روی همه آیتم ها به جز آیتم i هستند و $\text{sim}_{u,n}$ همان شباهت میان دو کاربر u و n است.

۳ سیستم های پیشنهادگر در گردشگری الکترونیکی

یکی از مهم ترین عوامل موفقیت یک سیستم پیشنهادگر گردشگری الکترونیکی و افزایش رضایتمندی گردشگران، هوشمندی این سیستم است [۱۳] یک سیستم هوشمند می تواند 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 می شوند. این نوع هوشمندی می تواند با الگوریتم های هوش مصنوعی و روش های 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 می رسد که کشورهای مختلف که دارای پتانسیل های گردشگری هستند در

قالب طرح‌های تحقیقاتی و کاربردی به طراحی و پیاده‌سازی ساختارهای تورسیم الکترونیکی پردازند 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 برقرار کنندگان 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 سرویس‌های موردنیاز گردشگران ساده‌تر، با کیفیتی بالاتر و هزینه‌ای 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 سرویس‌هایی که به واسطه فن آوری اطلاعات میسر شده‌اند [۱۷].

دلیل اینکه چرا کاربرد موفق تکنیک‌های پیشنهاد فیلم یا کتاب نمی‌توانند مستقیماً در حوزه جهانگردی بکار برده شوند این است که مدل کردن پروفایل‌های کاربری تورسیم 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 پیچیده‌تر از ساختار یک فیلم یا یک کتاب است. از طرفی تکنیک‌های 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 پیچیده‌تری داشته باشند در نتیجه ایجاد پروفایل‌های 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 سرویس‌های پیشنهادگر مسافرتی وجود دارد و تیم‌های تحقیقاتی زیادی در این زمینه مشغول به فعالیت هستند [۲۰].

۴ روش پیشنهادی

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

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

۴-۱ چگونگی تأثیر گذاری فاکتور های مختلف در سیستم های پیشنهاد گر تورسم

تحلیل ها نشان می دهد 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].

۴-۱-۱ سازماندهی همسایه ها

در اولین مرحله با ارزیابی روابط ماتریس کاربر-آیتم باید زیر مجموعه ای از نزدیک ترین کاربران به کاربر هدف (شبه ترین کاربران به کار هدف) را جستجو کرد. در این راستا، نمایه هر جفت کاربر a و b ، به منظور ارزیابی میزان شباهت آن ها، با یکدیگر مقایسه می شوند اگر $sim_{a,b}$ بیانگر میزان شباهت a و b باشد مقدار آن معمولاً در بازه 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 می شود. در این مرحله برای به دست آوردن یک تابع شباهت دقیق تر، ضریب تاثیر فاکتورهای مختلف در نرخ گذاری را به دست می آوریم و به تابع شباهت اولیه که از ماتریس کاربر-آیتم به دست آمده اضافه می کنیم. برای اینکار از نرم افزار داده کاوی کلمنتاین استفاده کرده و با استفاده از الگوریتم های ساخت درخت موجود در این نرم افزار ضریب تاثیر فاکتورهای مختلف را به دست می آوریم.

۴-۲ مجموعه داده

مجموعه داده ما در این تحقیق شامل ۲۷۶ کاربر، ۴۱ شهر و ۳۸۳۶ رتبه کاربران می باشد که از طریق پرسش نامه جمع آوری شده است. در این مجموعه داده هر کاربر رتبه ای بین ۱ تا ۵ به شهرها داده است که رتبه بالاتر نشان دهنده علاقه مندی بیشتر است [۱۵].

جدول ۲. نمونه ای 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	.	albacete	۲
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	.	alicante	۲
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	.	almeria	۳
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	.	asturias	۲
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	.	Avila	۲
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	.	badajoz	۳
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	.	baleares	۳
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	.	barcelona	۳
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	.	basque	۳
۱	۷۲۹۵۳۳/۵۱۷	۱۴۵۹۷۹۱۵/۷	۲۰	yes	Male	student	.	burgos	۴

user ID	coor X	coor Y	age	married	Gender	Occupation	Children	city name	rating
۲	۷۲۹۳۹۴/۴۱۶	۱۴۵۹۷۷۸۶/۹۴	۲۸	yes	Male	scientist	۲	almeria	۲
۲	۷۲۹۳۹۴/۴۱۶	۱۴۵۹۷۷۸۶/۹۴	۲۸	yes	Male	scientist	۲	asturias	۳
۲	۷۲۹۳۹۴/۴۱۶	۱۴۵۹۷۷۸۶/۹۴	۲۸	yes	Male	scientist	۲	avila	۳
۲	۷۲۹۳۹۴/۴۱۶	۱۴۵۹۷۷۸۶/۹۴	۲۸	yes	Male	scientist	۲	badajoz	۲
۲	۷۲۹۳۹۴/۴۱۶	۱۴۵۹۷۷۸۶/۹۴	۲۸	yes	Male	scientist	۲	baleares	۲
۲	۷۲۹۳۹۴/۴۱۶	۱۴۵۹۷۷۸۶/۹۴	۲۸	yes	Male	scientist	۲	barcelona	۲
۲	۷۲۹۳۹۴/۴۱۶	۱۴۵۹۷۷۸۶/۹۴	۲۸	yes	Male	scientist	۲	basque	۲
۲	۷۲۹۳۹۴/۴۱۶	۱۴۵۹۷۷۸۶/۹۴	۲۸	yes	Male	scientist	۲	burgos	۴
۲	۷۲۹۳۹۴/۴۱۶	۱۴۵۹۷۷۸۶/۹۴	۲۸	yes	Male	scientist	۲	caceres	۴
۲	۷۲۹۳۹۴/۴۱۶	۱۴۵۹۷۷۸۶/۹۴	۲۸	yes	Male	scientist	۲	cadiz	۳
۲	۷۲۹۳۹۴/۴۱۶	۱۴۵۹۷۷۸۶/۹۴	۲۸	yes	Male	scientist	۲	cantabria	۲
۳	۷۲۹۱۵۹/۹۵۱	۱۴۵۹۷۷۵۳/۳۵	۶۰	no	Female	professor	۰	avila	۲
۳	۷۲۹۱۵۹/۹۵۱	۱۴۵۹۷۷۵۳/۳۵	۶۰	no	Female	professor	۰	badajoz	۳
۳	۷۲۹۱۵۹/۹۵۱	۱۴۵۹۷۷۵۳/۳۵	۶۰	no	Female	professor	۰	baleares	۴
۳	۷۲۹۱۵۹/۹۵۱	۱۴۵۹۷۷۵۳/۳۵	۶۰	no	Female	professor	۰	barcelona	۳
۳	۷۲۹۱۵۹/۹۵۱	۱۴۵۹۷۷۵۳/۳۵	۶۰	no	Female	professor	۰	basque	۳
۳	۷۲۹۱۵۹/۹۵۱	۱۴۵۹۷۷۵۳/۳۵	۶۰	no	Female	professor	۰	burgos	۴
۳	۷۲۹۱۵۹/۹۵۱	۱۴۵۹۷۷۵۳/۳۵	۶۰	no	Female	professor	۰	caceres	۴
۳	۷۲۹۱۵۹/۹۵۱	۱۴۵۹۷۷۵۳/۳۵	۶۰	no	Female	professor	۰	cadiz	۲
۳	۷۲۹۱۵۹/۹۵۱	۱۴۵۹۷۷۵۳/۳۵	۶۰	no	Female	professor	۰	cantabria	۲
۳	۷۲۹۱۵۹/۹۵۱	۱۴۵۹۷۷۵۳/۳۵	۶۰	no	Female	professor	۰	castello	۳
۳	۷۲۹۱۵۹/۹۵۱	۱۴۵۹۷۷۵۳/۳۵	۶۰	no	Female	professor	۰	ciudad real	۳

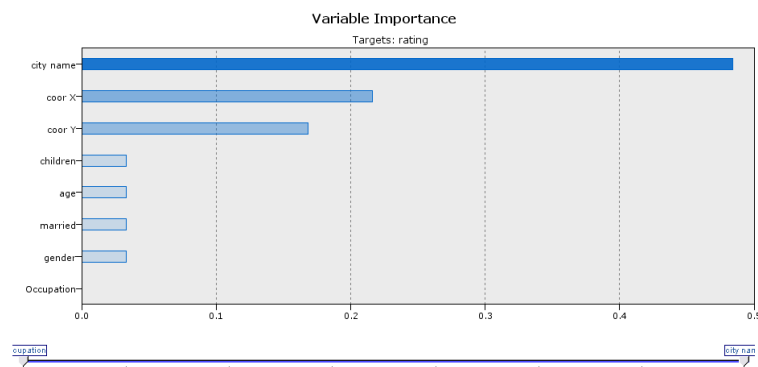
۴-۳ الگوریتم‌های ساخت درخت تصمیم

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 زیرگروه‌ها 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 رده‌ای باشند نتیجه کار یک درخت رده‌بندی خواهد بود.

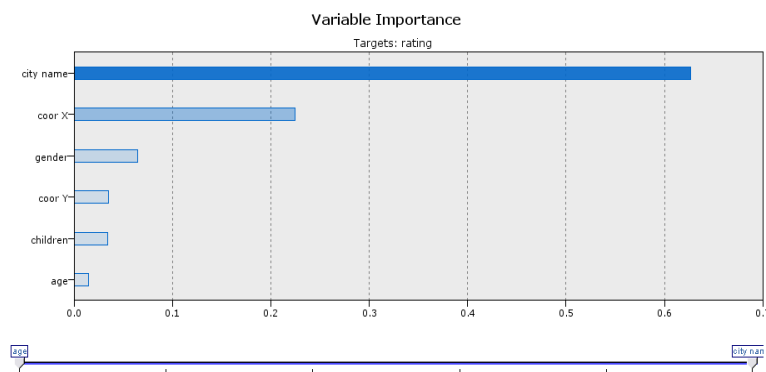
در مجموعه داده مورد استفاده در این تحقیق، چون فیلد هدف ما از نوع بازه‌ای است در نتیجه از بین ۴ مدل درختی موجود در نرم افزار کلمنتاین تنها ۲ مدل قابل استفاده بود، که عبارت‌اند از مدل C&R و مدل CHAID. درخت تقسیم و رگرسیون (C&R) یک درخت تصمیم تولید می‌کند که سعی در پیش‌بینی رده‌بندی مشاهدات آینده دارد. 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 پیش‌بینی‌کننده 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 می تواند 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 پیش بینی کننده و هدف می توانند هم از نوع بازه ای و هم از نوع رده ای باشد.

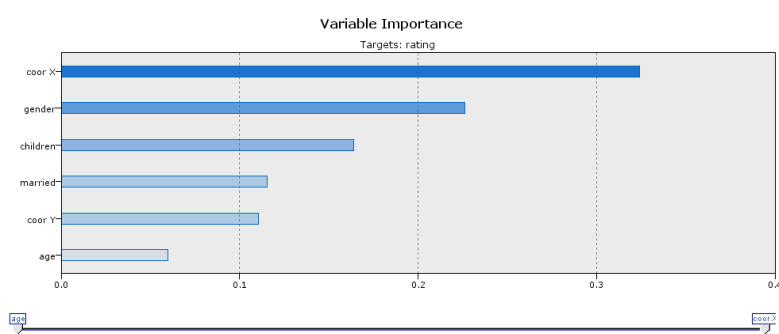
دو مدل دیگر درخت تصمیم موجود در نرم افزار کلمنتاین را نتوانستیم مورد استفاده قرار دهیم چون فیلد هدف در آن ها باید حتما از نوع رده ای باشد در حالیکه در مجموعه داده ما فیلد هدف از نوع بازه ای است. مجموعه داده مورد نظر توسط سه روش، مورد داده کاوی قرار گرفت که نتایج حاصل از آن ها در شکل های زیر قابل مشاهده است. در سه روش مورد آزمایش، مشاهده می شود که ضرایب تاثیر فاکتورها در مدل های 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 به دست 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 ۴. نتایج حاصل از آنالیز سه مدل

نتایج حاصل از آنالیز مدل ها نشان می دهد که مدل درخت تصمیم C&R خطای کمتری را در پیشگویی ها دارد. بنابراین از ضرایب به دست آمده از این مدل برای رسیدن به یک تابع شباهت جدید استفاده کرده و برای نرمال سازی این ضرایب، رابطه زیر را به کار می بریم:

$$\begin{aligned}
 &= \text{ضریب} \\
 &\left(\frac{\text{ضریب موقعیت مکانی}}{10^6} \times C \right) + \\
 &\left(\frac{\text{ضریب سن}}{1 + |\text{اختلاف سن دو کاربر}|} \times C \right) + \\
 &\left(C \times \text{تعداد شهرها با نرخ یکسان} \times \text{ضریب شهر} \right) + \\
 &\left(C \times \text{ضریب ازدواج} \right) : \text{در صورت یکسان بودن وضعیت ازدواج} + \\
 &\left(0 \times \text{ضریب ازدواج} \right) : \text{در غیر اینصورت} \\
 &+ \\
 &\left(C \times \text{ضریب جنسیت} \right) : \text{در صورت یکسان بودن جنسیت} + \\
 &\left(0 \times \text{ضریب جنسیت} \right) : \text{در غیر اینصورت} \\
 &+ \\
 &\left(C \times \text{ضریب شغل} \right) : \text{در صورت یکسان بودن شغل} + \\
 &\left(0 \times \text{ضریب شغل} \right) : \text{در غیر اینصورت}
 \end{aligned}
 \tag{۳}$$

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

رابطه بالا را برای هر دو کاربر محاسبه کرده و ماتریس جدیدی به دست می آوریم. سپس این ماتریس را به ماتریس شباهت اولیه اضافه نموده در نتیجه ماتریس شباهت جدید به دست می آید و در رابطه (۲-۲)، ماتریس شباهت جدید را به جای ماتریس شباهت اولیه قرار می دهیم که باعث می شود مرز همسایگی ها تغییر کرده و کاربرانی که به کاربر هدف شبیه ترند به عنوان همسایه نزدیک تر 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 می توان شباهت را از طریق ماتریس مشخصات کاربر به دست آورد.

در این بررسی ۸۰ درصد داده ها 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 مورد نظر، توسط درخت تصمیم C&R کمتر از دو روش دیگر است. بنابراین ضرایب به دست 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 می کنیم. برای اینکار ابتدا از روی ماتریس نرخ گذاری، شباهت میان کاربران را به دست می آوریم سپس با استفاده از تابع پیشگویی، رتبه شهرهایی که کاربر به آنها نرخ نداده را تخمین می زنیم و در نهایت از طریق تابع ارزیابی میزان خطا را محاسبه می کنیم.

جدول ۳. بخشی از ماتریس حاصل از تابع شباهت اولیه

	user 1	user 2	user 3	user 4	user 5
user 1	۱	۰/۱۴۷۴	۰/۸۴۰۲	۰	-۰/۱۲۵
user 2	۰/۱۴۷۴	۱	۰/۳۳۳	۰/۲۷۵۷	۰/۲۸۸۷
user 3	۰/۸۴۰۲	۰/۳۳۳	۱	۰/۵۲۳۸	۰/۱۶۶۷
user 4	۰	۰/۲۷۵۷	۰/۵۲۳۸	۱	۰
user 5	-۰/۱۲۵	۰/۲۸۸۷	۰/۱۶۶۷	۰	۱

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

جدول ۴. بخشی از ماتریس حاصل از تابع شباهت جدید

	user 1	user 2	user 3	user 4	user 5
user 1	۲	۱/۰۰۴۳	۱/۴۹۰۳	۰/۷۶۱۶	۰/۶۲۴۳
user 2	۱/۰۰۴۳	۲	۰/۹۸۴۱	۱/۰۳۰۲	۱/۰۳۸۶
user 3	۱/۴۹۰۱	۰/۹۸۴۱	۲	۱/۲۷۵۳	۱/۰۱۲۵
user 4	۰/۵۷۱۶	۱/۰۳۰۲	۱/۲۷۵۳	۲	۰/۶۵۲۶
user 5	۰/۶۲۴۳	۱/۰۳۸۶	۱/۰۱۲۵	۰/۶۵۲۶	۲

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

جدول ۶. 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 الگوریتم ها از برنامه MATLAB7.5 و برای اعمال روش های داده کاوی از نرم افزار SPSS Clementine12.0 استفاده شده است.

۵ ارزیابی مدل پیشنهادی

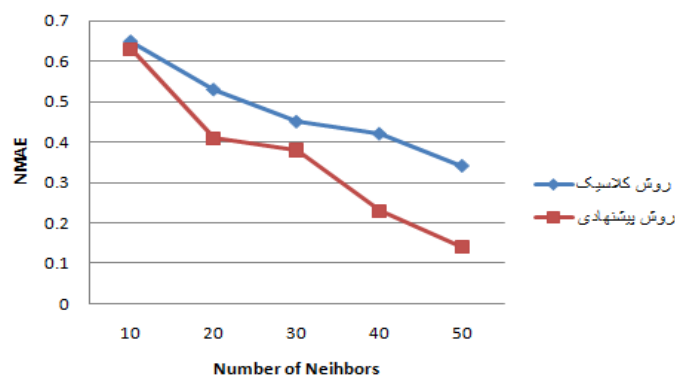
به منظور ارزیابی مدل پیشنهادی از معیار میانگین خطای مطلق نرمال شده (NMAE) استفاده می کنیم که پرکاربردترین معیار ارزیابی کیفیت در سیستم های پیشنهادگر می باشد. این معیار میانگین خطای میان پیشنهادات و رتبه های واقعی را برای یک مجموعه تست انتخاب شده محاسبه می کند [۳].

$$MAE = \frac{\sum_{(u,i) \in test} |prediction_{u,i} - real_{u,i}|}{n_{test}} \quad (4)$$

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

$$NMAE = \frac{MAE}{r_{max} - r_{min}} \quad (5)$$

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



شکل ۵. مقایسه روش پیشنهادی با روش کلاسیک

نمودار شکل ۵ رابطه بین NMAE با تعداد همسایه ها در روش پیشنهادی را با روش کلاسیک مقایسه می کند. همان طور که از نمودار مشخص است هرچه قدر تعداد همسایه ها بیشتر می شود NMAE در هر دو روش کاهش می یابد ولی در روش پیشنهادی، با به دست آوردن همسایه های بهتر برای کاربر هدف، توانستیم کیفیت پیشنهاد را بالاتر ببریم.

۶ نتیجه گیری

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

روش پیشنهادی ما در دو فاز اندازه گیری شباهت و آرایه پیشنهادات، الگوریتم های کلاسیک پالایش مشارکتی را اصلاح کرده است. در راستای محاسبه شباهت، افراد از جنبه های مختلفی با یکدیگر مقایسه شده اند 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 روش های 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 می توان همسایه های بهتری را برای کاربر هدف به دست 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.