

**یک روش مبتنی بر خوشه بندی برای بهبود کارایی در سیستم های توصیه گر مشارکت محور**

**چکیده**

در سیستم‏های توصیه‏گر مشارکت محور، محصولات به عنوان ویژگی شناخته می‎شوند و از کاربران درخواست می‌گردد تا به محصولات خریداری شده، رأی دهند. با آموختن رتبه‌بندی ارائه شده توسط کاربران، سیستم توصیه گر می‌تواند محصولات جالبی را به کاربران توصیه نماید. با این حال، معمولاً محصولات بسیار زیادی در تجارت الکترونیک وجود دارد و ممکن است این راه‌کار چندان کارآمد نباشد، به خصوص در زمان‌هایی که لازم است هر محصول قبل از ساخت در سیستم توصیه گر قرار گیرد. ما یک روش جدید را پیشنهاد می‌کنیم که یک الگوریتم خوشه‌بندی خودکار را برای کاهش ابعاد مرتبط با تعداد محصولات در بردارد. محصولات مشابه در خوشه‌ای یکسان قرار می‌گیرند و محصولات نامشابه، در خوشه‌های متفاوتی قرار می‌گیرند. کارهای پیشنهادی سپس با نتیجه خوشه‌بندی انجام می‌گیرد. در نهایت، انتقال مجدد انجام‌شده و لیست مرتب شده از محصولات پیشنهادی به هر کاربر پیشنهاد می‌گردد. با روش پیشنهادی، زمان پردازش برای تهیه پیشنهاد کاهش می‌یابد. نتایج آزمایش‌ها، نشان می‌دهد که کارایی سیستم های پیشنهادی می‌تواند، بدون به خطر انداختن کیفیت توصیه‌ها بهبود یابد.

**کلمات کلیدی:** سیستم توصیه کننده فیلتر مشارکتی. نمودار همبستگی; خوشه سازی خودساز. کاهش ابعاد؛ الگوریتم رتبه بندی

**1. مقدمه**

با توجه به توسعه سریع تجارت الکترونیکی، امروزه تعداد زیادی خریدار و فروشنده آنلاین وجود دارد و مقدار زیادی محصولات وجود دارد که کاربران می‌توانند برای خرید انتخاب کنند. به هر حال، وظیفه بررسی و انتخاب محصولات مناسب از جمله تعداد زیادی از محصولات، نه تنها گیج‌کننده، بلکه وقت‌گیر است. سیستم های توصیه گر [1،2] برای کمک به مردم در یافتن محصولات جالب‌توجه و ذخیره زمان جستجوی محصولات کمک می‌کنند. برای یک کاربر، چنین سیستمی می‌تواند از تجربیات ذخیره‌شده برای تمام مشتری‌ها و توصیه‌های یک لیست اولویت از محصولات به کاربران استفاده کنند. در طول چند سال گذشته، سیستم های پیشنهاددهنده به صرعت در حال تحول هستند. بسیاری از سیستم های توصیه گر توسعه یافته‌اند.

در اصل می‎توان آن‌ها را در دو دسته طبقه‌بندی کرد: محتوا محور و مشارکت محور. اگر چه گرایش به سمت سیستم های ترکیبی [3] در سال‌های اخیر افزایش یافته است. سیستم توصیه گر مبتنی بر محتوا [4] به کاربر بر اساس محتوایی که ممکن است شامل دسته‌ها و یا ویژگی‌های دیگر محصولات باشد، پیشنهاد می‌دهد. همچنین ممکن است به عادات، منافع و یا تنظیمات کاربران مراجعه کند. با تجزیه و تحلیل این داده‌ها با برخی از فناوری‌ها از جمله مدل‌سازی بیزین [8،9]، محتوای مبتنی بر سیستم پیشنهاددهنده جذاب تر خواهد بود. به طور کلی، سیستم های مبتنی بر محتوا نیاز به اطلاعات دقیق در مورد محصولات و کاربران ندارند. محصولات جدید را نیز می‌توانند به کاربران توصیه نمایند. با این حال، رسیدن به اطلاعات مورد نیاز سخت و یا زیاد است. محصولات یا کاربران دارای ویژگی‌های خود هستند. مشکل جمع‌آوری ویژگی‌ها از تمام محصولات و کاربران در هر صورت وجود دارد. علاوه بر این، اطمینان از اینکه یک محصول یا یک کاربر، را می‌توان با داده‌های جمع‌آوری شده نمایش داد، مشکل است. یک سیستم توصیه گر مشارکت محور، نیازی به جزئیات اطلاعات درباره ویژگی‌های محصولات یا کاربران ندارد. به جای آن، آن با اجرای تعاملات بین اطلاعات کاربران و محصولات توصیه می‌کند. همواره، اطلاعات تعامل به عنوان رتبه‌بندی کاربر برای محصولات خریداری شده بیان می‌شود. با یادگیری این رتبه‌بندی، چنین سیستم پیشنهاددهنده‌ای، می‌تواند محصولات را بر اساس نظرات ثبت‌شده سایر کاربران و ترجیحات خود کاربر مدنظر انجام دهد. به طور کلی، سیستم های توصیه گر مشارکت محور، ساده و موثر هستند و در جامعه تجارت، جذاب تر و عملی تر هستند.

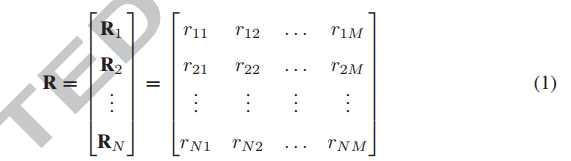
معمولاً محصولات بسیار زیادی برای بررسی در سیستم های توصیه گر وجود دارند و این امر می‌تواند بسیار ناکارآمد باشد، اگر هر محصول نیاز داشته باشد قبل از دریافت توصیه بررسی گردد. روش کاهش ابعاد برای تولید سریع توصیه باکیفیت بالا و مجموعه‌هایی با مقیاس بزرگ استفاده می‌گردد. در [26]، یک روش K-means به نام الگوریتم خوشه‌بندی K-means برای گروه‌بندی کاربران درون گروه‌های متفاوت ارائه شده است. همسایگان هر کاربر به دنبال تقسیم‏بندی کاربران، برای دریافت توصیه از سیستم توصیه گر از نتایج سایر کاربران هم گروه با خود بهره می‌گیرند. ژو و همکارانش [27] یک رویکرد مبتنی بر Smooth ارائه کردند که از خوشه‌های تولید شده از داده‌های آموزشی برای تهیه خوشه‌ها و همسایگان هر خوشه بهره می‌گیرد. Honda و همکارانش [28]، یک روش خوشه‌بندی برای انتخاب همسایگان بر اساس یک نظریه تعادل ساختاری ارائه کردند. کاربران و محصولات بر اساس متعادل کردن گراف ترکیبی، به خوشه‌هایی تقسیم می‌شوند.

Be و همکارانش [29]، کاربران را در خوشه‌هایی مطابق با صفات آنان خوشه‌بندی کردند، به عنوان مثال، بر اساس جنسیت، سن و شغل. سپس ماتریس امتیاز کاربر به هر کالا را تجزیه و تحلیل کرده و شباهت کاربران به یکدیگر را محاسبه کردند. سیستم iExpand [30]، مشارکت محور بودن سیستم های توصیه گر را توسط علایق کاربران و رتبه‌بندی شخصی هر کاربر انجام داد. معرفی یک طرح سه لایه‌ای، برای کمک به فهم تعاملات بین کاربران، محصولات و کاربران توسط آنان انجام شد. LDA، برای تقسیم‌بندی محصولات به خوشه‌هایی ارائه گردید. در [31]، خوشه‌بندی فازی در کاربران، به تشکیل گروه‌های کاربران منجر شد. یک بردار خصوصیت گروهی برای هر کاربر ساخته می‌شود. هر بردار خصوصیت کاربر را می‌توان به عنوان یک بردار نمایش‌دهنده میزان اهمیت و کارایی محصولات برای آن کاربر در نظر گرفت. Sarwat و همکارانش [32]، سیستم توصیه گری را پیشنهاد دادند که از یک طبقه‌بندی کننده با سه نوع نرخ مبتنی بر مکان استفاده می‌کند. در PRM2 [33]، منابع شخصی، شباهت علایق فردی و نفوذ فردی تجمیع شده و توسط روش SVD، کاهش بعد یافته و مورد استفاده قرار می‌گیرند. BiFu [34] یک مفهوم از محصولات محبوب را شناسایی کرده و برای توصیه در اختیار سیستم های توصیه گر قرارداد.

برای کاهش ابعاد ماتریس رتبه‌بندی، K-means [35] برای گروه‌بندی کاربران و محصولات درون خوشه‌هایی استفاده شده است. همچنین از روش smoothing and fusion برای غلبه بر پراکندگی داده‌ها و تنوع امتیاز استفاده می‌گردد. ICRRS [36]، یک الگوریتم تکراری است که نه تنها مبتنی بر ارزیابی، بلکه مبتنی بر مقایسه و تقریب زنی است. سیستم های توصیه گر مبتنی بر کاهش ابعاد ذکر شده در بالا، دارای برخی معایب هستند. برخی از سیستم ها [29، 32]، نیاز به ویژگی‌های اضافی دیگری در مورد کاربران و یا محصولات دارند. این ویژگی‌ها معمولاً برای دستیابی به کاربرد عملی ضروری است. به عنوان مثال، [26]، [27] و ][31]، نیاز به خوشه‌های بیشتری برای پیشبرد هدف دارند. همچنین شباهت اندازه‌گیری شده، اکثراً برای سیستم های اتخاذ شده در زمینه کاهش ابعاد مورد استفاده قرار می‌گیرد. نادیده گرفتن واریانس خوشه‌بندی، ممکن است منجر به نتایج نادرستی گردد. در این مقاله، یک رویکرد مبتنی بر خوشه‌بندی خودساخته برای اعمال پیشنهاد با هدف کاهش ابعاد مرتبط با تعداد محصولات ارائه شده است. محصولات مشابه در هر خوشه‌بندی، برای انجام توصیه در همان خوشه مورد استفاده قرار می‌گیرند. نمودار همبستگی نشان می‌دهد که ارتباط متقابل میان گروه محصولات ایجادشده ضروری و مهم است. الگوریتم random walk، اجراشده و لیست اولویت‌های محصولات، گروه‌بندی شده و مورد استفاده قرار میگرند. پس از آن، لیست اولویت، مجدداً به لیست اولویت‌های فردی از محصولات تبدیل می‌گردد. در نهایت، با روش پیشنهادی، خوشه‌بندی کاملا بر اساس ماتریس اولویت دهی بدون نیاز به جمع‌آوری صفات اضافی درباره محصولات و کاربران انجام می‌گیرد. خوشه‌ها به صورت خودکار تولید شده و یک تعداد خوشه از پیش تعیین شده توسط کاربر ایجاد می‌گردد. علاوه بر این، در هنگام نیاز به اندازه‌گیری شباهت برای خوشه‌بندی، ما از واریانس مراکز خوشه‌ها استفاده می‌کنیم. با توجه به گاهش ابعاد در تعداد محصولات، زمان پردازش برای ساخت توصیه توسط رویکرد ما بسیار کاهش می‌یابد. نتایج تجربی نشان می‌دهد که بهره وری سیستم پیشنهادی را می‌توان بدون کاهش کیفیت، تا حد زیادی بالا برد. ادامه مقاله به شرح زیر سازمان‌دهی شده است. مسئله‌ای که باید حل شود، در بخش 2 نشان داده شده است. یک سیستم توصیه گر مشارکت محور، ItemRank [17]، در بخش 3 معرفی شده است. روش پیشنهادی ما برای بهبود کارایی ItemRank، با جزئیات در بخش 4 ارائه شده است. یک مثال در بخش 5 نشان داده شده است. نتایج آزمایش‌ها در بخش 6 نمایش داده شده است. در نهایت، بخش پایانی، نتیجه‌گیری را نشان می‌دهد.

**2. بیان مسئله**

فرض کنید یک مجموعه از N کاربر به صورت 1<i<N وجود دارد و یک مجموعه از M محصول pj، 1<j<M وجود دارد. یک کاربر ui، ارزیابی خود از محصولات pj را با تهیه یک نرخ rij نمایش می‌دهد. همواره، یک مقدار بزرگ‌تر برای نمایش یک بازخورد از کاربر شروری است. اگر کاربر ui، اولویت و یا امتیازی را برای محصول pj وارد نکند، rij=0. چنین اطلاعاتی را می‌توان توسط ماتریس R زیر نمایش داد.



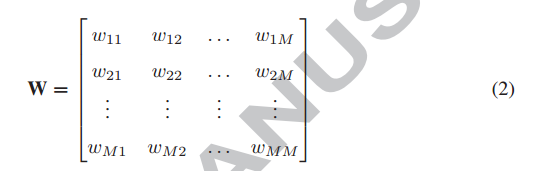
که یک ماتریس N در M است. توجه داشته باشید که

.

برای راحتی، هر سطر، یک رکورد کاربر نامیده می‌شود و هر ستون، نشانگر یک محصول است. هدف از سیستم توصیه گر مشارکت محور، رتبه‌بندی ماتریس داده است که در آن یک لیست اولویت پیش‌بینی‌شده از محصولات به هر کاربر تهیه می‌شود.

**3. ItemRank**

ItemRank [17] یکی از روش‌های اساسی برای مشارکت بودن سیستم های توصیه گر است. آن ِک الگوریتم امتیازدهی مبتنی بر random-walk، برای پیشنهاد محصولات مطابق با اولویت کاربر استفاده می‌کند. ItemRank در مطالعات ما استفاده شده است، زیرا دارای پیچیدگی کمتر و کارایی بالاتر در مصرف پول و هزینه محاسباتی نسبت به سایر سیستمها است . ماتریس اولویت نمایش داده شده در رابطه (1)، دارای فرآیند ItemRank با دو گام است، ایجاد گراف همبستگی و random walk [38]. در گام ایجاد گراف همبستگی، یک گراف همبستگی ساخته می‌شود. هر محصولی به عنوان یک محصول در داخل گراف است و یک گراف همبستگی از اولویت‌های داده شده ساخته می‌شود. هر محصولی، به عنوان یک نود در گراف است. یال بین هر دو نود، pi و pj، دارای وزن wij است که نشانگر تعداد کاربرانی است که به هر دو محصول، امتیاز یکسانی داده بودند. توجه داشته باشند یک کاربر uk، اولویت هر دو محصول را تعیین می‌کند. زمانی که گراف همبستگی کامل شد، ماتریس همبستگی زیر ایجاد می‌گردد:



که یک ماتریس M در M است. هر ستون W، نرمال سازی می‌گردد. در مرحله random walk، یک سری از random walk ها اجرا می‌گردد. برای بررسی وضعیت هر کاربر، Si(0) را داریم:



که برداری با سایز M است. عملیات زیر



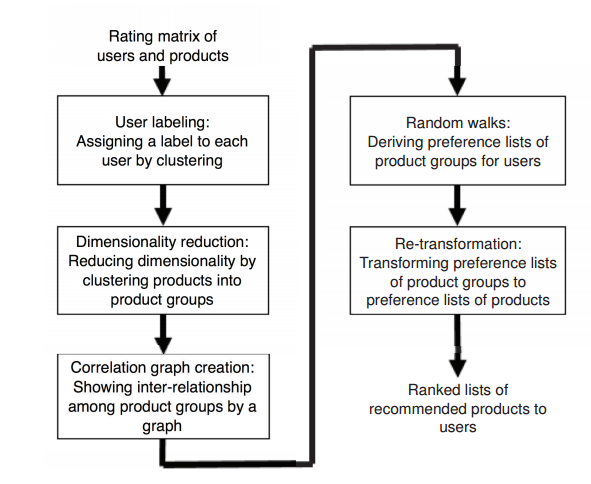
مکرراً برای t=0,1,2,… اجرا می‌گردد تا همگرایی ایجاد گردد. توجه داشته باشید که . همواره، یک همگرایی قابل‌قبول، بعد از تقریباً 20 تکرار ایجاد می‌گردد. بنابراین، ضروری است که رابطه (4) را حد.د 20 بار برای هر کاربر اجرا کنیم. اگر Si بردار همگرایی باشد، اولویت پیشگویی محصولات کاربر ui قابل محاسبه است. سپس می‌توان محصولاتی را برای کاربر ui توصیه کرد. محصولات متعلق به بزرگ‌ترین عنصر در Si، اولین توصیه است و دومین عنصر بزرگ در Si، دومین توصیه است.

**4. روش پیشنهادی**

Itemrank مسئله کارایی را مدنظر قرار می‌دهد. زیرا ممکن است تعداد زیادی محصولات در تجارت الکترونیک وجود داشته باشد. ماتریس W، که دارای سایز M\*M است، می‎تواند شدیداً بزرگ باشد. تکثیر W با Si(t) در رابطه (4)، زمان زیادی را می‌گیرد و روش ItemRank را برای مقیاس بزرگ مسئله ناکارآمد می‌کند. ما یک الگوریتم خوشه‌بندی خودساخته (SCC)، [39،40] اجرا کرده‌ایم که عمل کاهش بعد را برای تولید خوشه‌ها استفاده می‌کند. کار پیشنهادی سپس با خوشه‌بندی مجدد ادامه می‌یابد. در نتیجه، کارایی ItemRank، می‌تواند بهبود یابد. در مقایسه با سایر روش‌های کاهش بعد [41، 42، 43، 44]، SCC دارای مزایای زیادی است. خوشه‌بندی به صورت اتوماتیک انجام می‌گیرد و تعیین تعداد خوشه‏ها توسط کاربر، مورد نیاز نیست. در کنار آن، زمانی که شباهت خوشه‌ها اندازه‌گیری گردید، هر دوی مراکز و واریانس خوشه‌ها محاسبه می‌گردند. در پایان، اندازه‌گیری شباهت بهتر از بررسی مراکز در سایر روش‌ها است.

روش ما شامل پنج گام است. برچسب‌گذاری کاربر، کاهش بعد، ایجاد گراف همبستگی، random walk و انتقال مجدد که در شکل 1 نشان داده شده است. در کام برچسب‌گذاری کاربر، SCC، برای نگاشت برچسب کلاس به کاربر برای کمک به گام دوم انجام می‌گیرد تا مرحله دوم بتواند کار خود را به طور کارا انجام دهد. در گام کاهش بعد، SCC، مجدداً برای خوشه‌بندی تعدادی از گروه محصولات استفاده می‌گردد. محصولات مشابه، به گروه محصولات یکسانی تعلق دارند و محصولات غیرمشابه، به محصولات متفاوتی تعلق دارند. از آنجایی که تعداد گروه محصولات کوچک‌تر از تعداد محصولات است، ابعاد درگیر کاهش خواهد یافت. سپس یک گراف همبستگی که نشانگر رابطه داخلی بین گروه محصولات هست، ایجاد می‌گردد. بر اساس گراف همبستگی، یک سری از random walk ها اجراشده و یک لیست اولویت از گروه محصولات برای هر کاربر در گام چهارم ایجاد می‌گردد. در نهایت، در گام انتقال مجدد، لیست اولویت گروه محصولات به لیست اولویت تکی محصولات منتقل می‌گردد و یک لیست رتبه‌بندی شده از محصولات برای هر کاربر ایجاد می‌گردد.

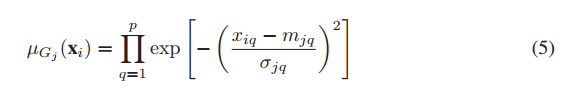
در این مطالعه، ما ItemRank [17] را به عنوان هدف نهایی استفاده کرده‏ایم. به هر حال، روش ما می‌تواند با سایر روش‌ها برای کاهش ابعاد و بهبود کارایی آن‌ها کار کند.



شکل 1. بررسی اجمالی از رویکرد ما.

* 1. **خوشه‌بندی خودساخته (SCC)**

مجموعه X با n پارامتر x1,x2,…,xn با  برای 1<i<n، نشان داده شده است. هدف الگوریتم SCC، گروه‌بندی این الگوها درون یک مجموعه از کلاسترها است، تا الگوهای یکسان در یک گروه و الگوهای غیریکسان درون گروه‌های مختلفی قرار گیرند. اگر K، تعداد خوشه‌های جاری موجود با نام  باشد. هر خوشه Gj، دارای میانه  و واریانس  است که حاصل میانگین و واریانس تمام الگوهای موجود است. اگر Sj، سایز کلاستر Gj باشد، تعداد الگوهای شامل Gj خواهد بود. بنابراین ما داریم K=0 و این مورد نشانگر عدم وجود هیچ خوشه‌ای است. برای هر الگوی xi، 1<i<n، ما اعضای آن با درجه xi را محاسبه کرده‌ایم که درون خوشه  وجود دارد.





که در آن  یک حد آستانه است. توجه داشته باشید که علامت "رو" نقش مهمی را در این الگوریتم خوشه‌بندی بازی می‌کند. مقدار بزرگ برای علامت "رو" در خوشه‌های کوچک‌تر استفاده می‌شود، اگر چه مقدار کوچک آن در خوشه‌های بزرگ است. همان طور که "رو" افزایش می‌یابد، تعداد خوشه‌ها نیز افزایش می‌یابد. دو مورد ممکن است رخ دهد. اولاً، هیچ خوشه‌ای بر روی xi نباشد. در این مورد، ما فرض می‌کنیم که xi، مشابه با سایر خوشه‌ها نیست و خوشه جدیدی با h=K+1 ایجاد می‌گردد:



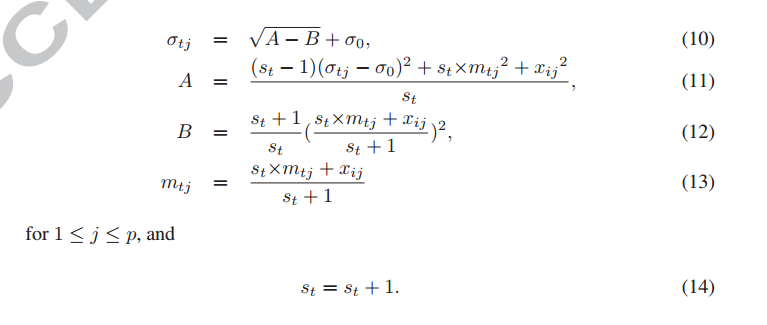
به طوری که در آن  یک بردار ثابت تعریف شده توسط کاربر است. البته، تعداد خوشه‌ها افزایش می‌یابد و سایز خوشه Gh و Sh باید برابر با یک باشد.



دوما، خوشه‌هایی که در Xi وجود دارد، دارای تست مشابهی هستند. در این مورد، خوشه Gt خوشه‌ای با بزرگ‌ترین اعضا است.



Xi باید مشابه با خوشه Gt باشد و mt و Gt باید تغییر یابند. تغییرات خوشه Gt به صورت زیر هستند:

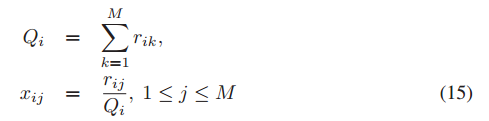


توجه داشته باشید که K، در این مورد تغییری نمی‌کند. این فرایند تا زمانی که الگوها قابل‌اجرا باشند، ادامه می‌یابد. مطابق با آن، ما K خوشه برای X محاسبه کرده‌ایم.

* 1. **گام 1: برچسب‌گذاری کاربر**

برای کاهش ابعاد به طور موثر، نیاز به تعریف برچسب کلاس برای کاربران داریم. ایده، گروه‌بندی کاربران درون خوشه‌هایی است [32]. کاربران مشابه، درون خوشه‌هایی گروه‌بندی شده و کاربران غیرمشابه، درون خوشه‌های مختلفی قرار می‌گیرند. سپس تمام کاربران در یک گروه، برچسب یکسانی را دریافت می‌کنند. ما از الگوریتم SCC برای این هدف استفاده کردیم. الگوریتم‌های خوشه‌بندی دیگری [45،46،35] نیز می‌توانند این کار را انجام دهند. اما، آن‌ها نیاز دارند تا ای پیش در مورد تعداد کلاس‌ها تصمیم‌گیری کنند. با الگوریتم SCC، ما فقط نیاز داریم تا برخی ثوابت معنادار را در طول فرآیند خوشه ایجاد کنیم.

برای اعمال SCC، ما شباهت ایجادشده میان کاربران را به صورت رتبه‌بندی شده ایجاد کردیم. با این حال، مردم دارای شخصیت های مختلفی هستند و برای رأی دادن به محصولات از سلایق مختلف خود استفاده می‌کنند. در واقع برخی سخاوتمند بوده و نمرات بیشتری را در نظر می‌گیرند و در مقابل برخی سخاوتمند نیستند و نمرات کمتری را در نظر می‌گیرند. حال می‌توان با توجه به رتبه‌بندی یک کاربر، برای دیگران توصیه در نظر گرفت. فرض کنید که دو کاربر شبیه به یکدیگر هستند و شکل موج ایجادشده توسط رتبه‌بندی آن‌ها یکسان است. بنابراین ما برای رتبه دهی به کاربران از رابطه زیر استفاده می‌کنیم:



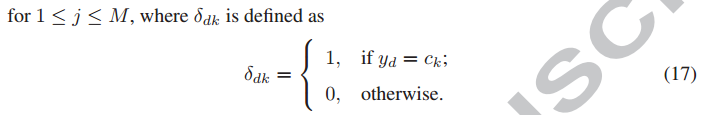
برای 1<i<N، داریم، 

ما الگوریتم SCC را بر روی X اعمال می‌کنیم. فرض کنید که z کلاستر G1,G2,…,Gz وجود داشته باشد. هر کلاستر به عنوان یک کلاس است و ما z کلاس داریم که دارای برچسب‌های c1,c2,…,cz هستند. برای تمام کلاسترها در Gj، 1<j<z. برچسب کلاس cj را داریم. در نتیجه، ما مجموعه R را با مجموعه دیگر R’ که شامل N مؤلفه است ایجاد کرده‌ایم: 

* 1. **کاهش ابعاد**

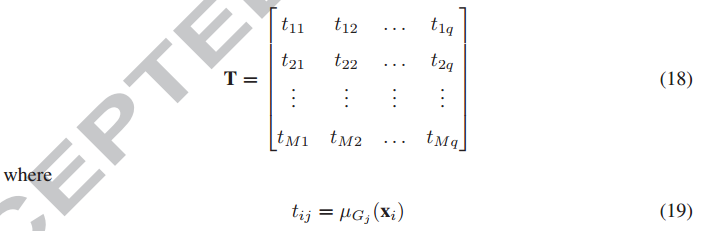
در این گام، ما ابعاد M را کاهش دادیم که مرتبط با محصولات مشابه بود. برای هر محصول pj، 1<j<M است. ما یک الگوی ویژگی ایجاد کردیم. 



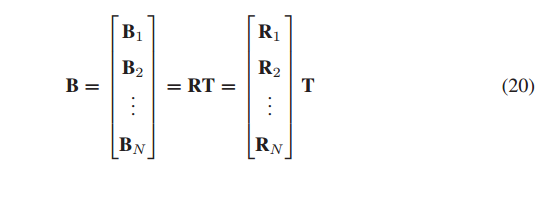


بنابراین، ما ویژگی x1,x2,…,xM را برای هر مؤلفه z ایجاد کردیم. داریم 

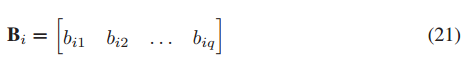
در نهایت، ما الگوریتم SCC را بر روی Y اجرا کردیم. فرض کنید که q خوشه ایجاد کرده‌ایم. توجه داشته باشید که محصولات شامل خوشه‌های مشابه هستند. این قابل‌قبول است که خوشه‌ها مسئول تمام محصولات هستند. از آنجا که در آن، q کلاستر وجود دارد، یک کاربر باید M مؤلفه را انتخاب کند. ما می‌توانیم ابعاد M را کاهش داده و یک q با ابعاد کوچک‌تر داشته باشیم. T ماتریس کاهش بعد یافته است:



که درجه عضویت xi را در خوشه Gj در رابطه (5) نمایش می‌دهد. سپس ما R کاهش بعد یافته را درون یک ماتریس N\*M قراردادیم تا



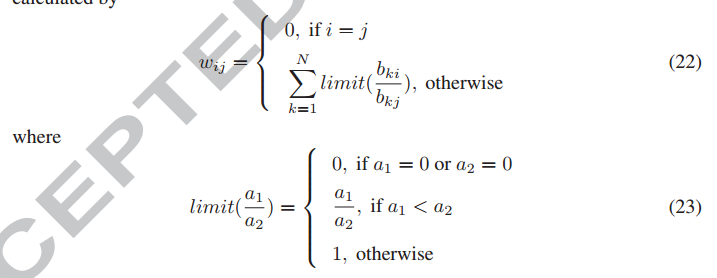
که در آن B یک ماتریس N\*q است. توجه داشته باشید که



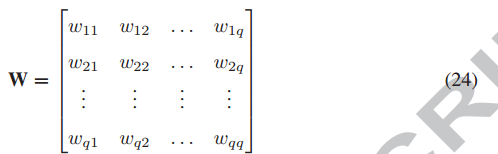
برای سهولت، ما ستون B را به عنوان یک گروه محصولات در نظر گرفته‌ایم. بنابراین، ما q گروه از محصولات را داریم که با نام  به ترتیب نشان داده شده‌اند. در این راستا، کاربر باید M مؤلفه را ذخیره کند و هر مؤلفه متعلق به یک محصول است. Itemrank با ماتریس R کار می‌کند. اما روش ما، با ماتریس B کار می‌کند. از آنجایی که q، همواره از M بزرگ‌تر است، ابعاد می‌توانند کاهش‌یافته و به طور چشمگیری بهبود یابند.

**4.4 گام 3: ایجاد گراف همبستگی**

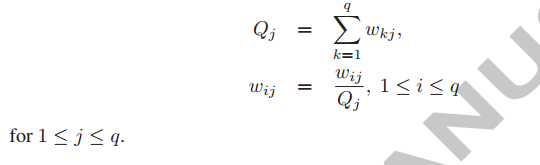
در این گام، ما یک گراف همبستگی ایجاد کرده‌ایم که نشانگر روابط بین محصولات گروهی q است [47]. چون ما با B کار می‌کنیم، گراف ما دارای تفاوت‌هایی خواهد بود. هر گروه از محصولات، به عنوان یک گره هستند و بنابراین، ما q گره در گراف داریم. وزن wij بین گره gi با 1<i و J<q، به صورت زیر تعریف می‌گردد:



برای موردی که ، داریم: . زیرا دارای اولویت کمتری برای کاربر است. برای موردی که ، ما داریم،  تا کاربر را برای داشتن اولویت بهتر تشویق کنیم. به هر حال، اگر Wij خیلی بزرگ باشد، برخی از معیارهای گروهی، بر دیگری غلبه خواهند کرد. بنابراین ما یک محدودیت بر روی wij ایجاد می‌کنیم. زمانی که گراف همبستگی کامل شد، ما ماتریس همبستگی زیر را داریم:



که در آن، ماتریس، یک ماتریس q در q است. هر ستون از W، نرمال می‌گردد:



**4.5 گام چهارم: random walk**

در این گام، یک سری از الگوریتم‌های random walk [48،38،17،49] ارائه شده است



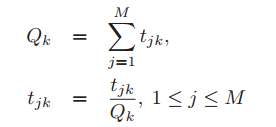
که دارای سایز برداری برابر با q است.



توجه داشته باشید که W یک ماتریس در رابطه (24) است و Bi، برداری در رابطه (21) است. همانند آنچه در [17] نشان داده شده است، آلفا، در رابطه (4) انتخاب‌شده تا برابر با 0.85 باشد. Vi، لیست اولویت ایجادشده برای سهولت گروه‌بندی است.

* 1. **گام پنجم: Re-Transmision**

هر بردار Vi، برای کاربر ui است. مقدار q، نشانگر گروه‌بندی محصولات است. به هر حال، ما نیاز داریم تا گروه‌بندی شخصی را برای هر کاربر رعایت کنیم. رابطه (19) دارای درجه xj است. بنابراین داریم:



برای است. برای هر سطر، این محاسبات شامل محصولات Pj است. بنابراین داریم:

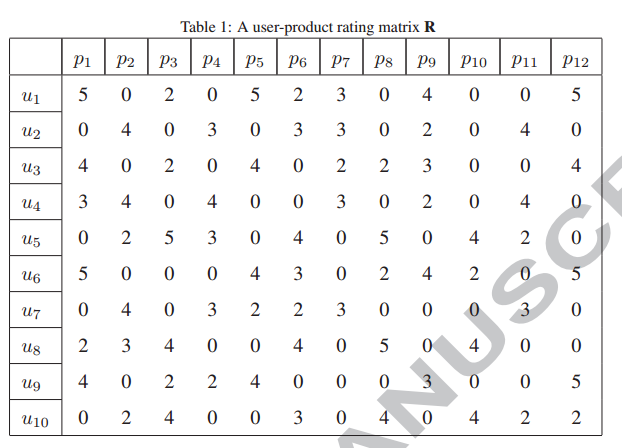


که در آن Si[j]، jامین مؤلفه Si است و Vi[k] در آن 1<k<q، iامین مؤلفه Vi است. توجه داشته باشید که tjk، محصول مشابه با محصولات pj است. بنابراین ، اولویت محصول pj را برای کاربر ui در مورد گروه محصولات gk نشان می‌دهد. در مجموع، نتایج در رابطه (27) نشان داده شده است. نتایج اولویت محصولات pj برای کاربر ui است. در نهایت، ما لیست اولویت را در لیست Si قراردادیم.

**4.7 آنالیز پیچیدگی**

ما آنالیز پیچیدگی محاسباتی راه‌کار پیشنهادی را در این بخش انجام داده‌ایم. در گام برچسب‌گذاری کاربر، ما شباهت بین کاربرها و خوشه‌های هر یک را محاسبه کردیم. بازگو می‌کنیم که N، تعداد کاربران را نشان می‌دهد و M، تعداد محصولات را نشان می‌دهد. پیچیدگی زمانی این گام برابر است با O(NzM). در گام کاهش ابعاد، ما شباهت بین الگوهای بار و خوشه‌های موجود را بررسی کردیم. زیرا تعداد الگوهای ویژگی برابر با M است و تعداد محصولات گروه‌بندی شده برابر با q است. در گراف همبستگی ایجادشده، هر وزن wij نشانگر رابطه (22) است.

جدول 1. یک ماتریس رتبه دهی R

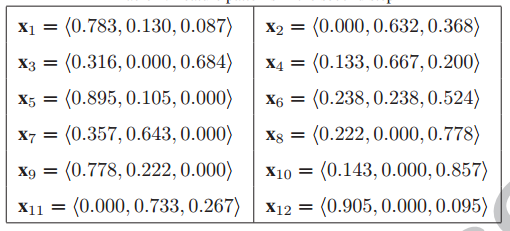


بنابراین، پیچیدگی این مرحله به صورت O(q2) است. در مرحله random walk، رابطه (26) دارای یک روند تکراری است. برای هر تکرار، WVi(t) محاسبه می‌گردد. زیرا 20 تکرار قبل از همگرایی لازم است. در نهایت، برای یک کاربر، رابطه (27)، دارای زمان محاسباتی M است. پیچیدگی این مرحله به صورت  است. بنابراین پیچیدگی زمانی راه‌کار ما برای تمام محصولات به صورت  است.

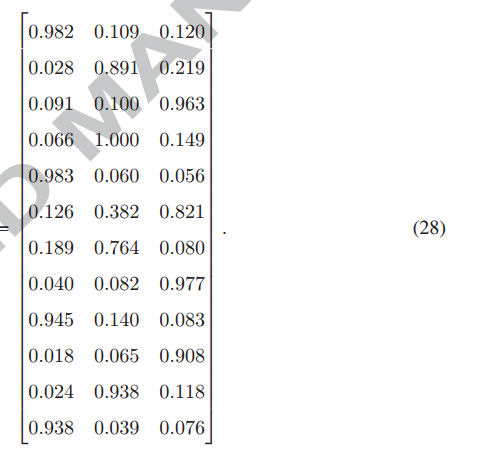
**5. مثال**

در این بخش، ما یک مثال برای نمایش راه‌کار پیشنهادی ارایه داده‌ایم. ماتریس R مورد بررسی در جدول 1 نشان داده شده است. توجه داشته باشید که 10 کاربر و N=10 , 12 محصول وجود دارد و M=12, 65 است. برای مثال، کاربر u1، دارای 7 نرخ تولید محصول، به صورت  است. در مرحله برچسب‌گذاری کاربر، برچسب c1، به کاربر  اعطا می‌گردد و برچسب کلاس c2، به کاربران  و برچسب c3 به کاربران  اعطا می‌گردد.

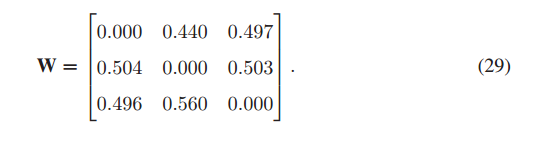
جدول 2. ویژگی الگوها در گام دوم



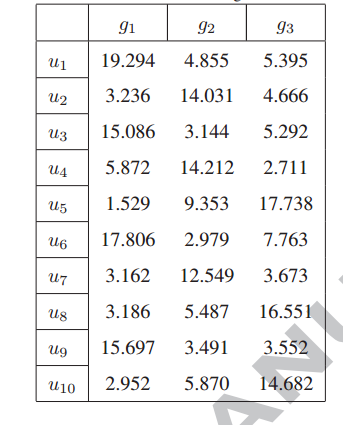
در مرحله کاهش ابعاد، ما الگوهای ویژگی را به صورت نمایش داده شده در جدول 2 در نظر گرفته‌ایم. پس از اجرای الگوریتم SCC، سه خوشه G1, G2, G3 به دست می‌آیند. ماتریس کاهش بعد یافته T به صورت زیر است:

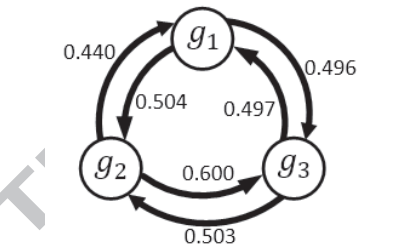


با اجرای رابطه (20)، ما ماتریس R را کاهش داده و به ماتریس B می‌رسیم که در جدول 3 نشان داده شده است. توجه داشته باشید که سه گروه محصول با عناوین g1, g2,g3 وجود دارد. ما گراف همبستگی ایجادشده در گام قبل را داریم. با اجرای رابطه (22)، ما گراف همبستگی را همانند آنچه در شکل 2 نشان داده شده است، داریم.



جدول 3. ماتریس کاهش بعد یافته B

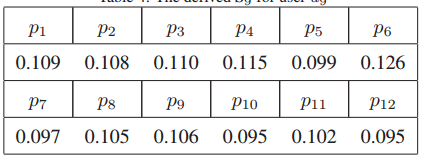




شکل 2. گراف همبستگی

در مرحله random walk، ما رابطه (26) را برای هر کاربر اجرا می‌کنیم. برای نمونه، پس از هر همگرایی، ما  را داریم. در نهایت، در گام re-transformation، ما Vi را به Si تبدیل کردیم. برای مثال، V9، به S9 برای کاربر u9 تبدیل می‌شود. مؤلفه‌های S9 در جدول 4 نشان داده شده است. چون p6، دارای مقدار اولویت بزرگ‌تری است، 0.126 در S9 و p­6­، اولین توصیه برای کاربر u9 و سپس p4 و p3 خواهد بود.

جدول 4. مشتق شده S9 برای کاربر U9

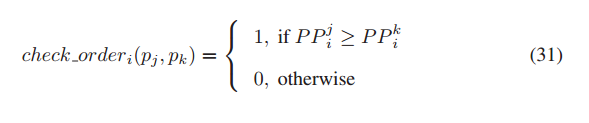


**6. نتایج آزمایش‌ها**

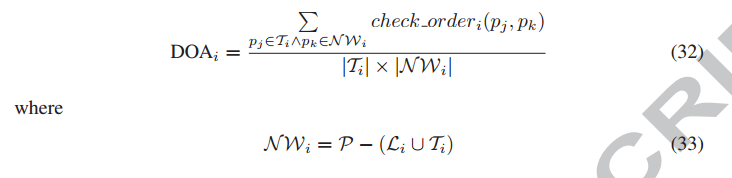
برای ارزیابی کارایی الگوریتم مبتنی خوشه‌بندی خودساخته (SCC)، ما یک مجموعه از آزمایش‌ها را بر روی چندین مجموعه داده انجام دادیم. برای همگرایی، ما روش SCC خود را در این بخش ارزیابی کردیم. ما همچنان، روش SCC را یا سایر راه‌کارهای مبتنی بر فیلترگذاری مقایسه کردیم. سه متریک برای مقایسه بر روی دقت توصیه‌ها ارائه شده است: درجه توافق (DOA)، میانگین خطا و خطای میانی ریشه. یک اعتبارسنجی برای آزمایش‌ها ارائه شده است. در این آزمایش‌ها، ورودی، مجموعه داده‌ها با 5 زیرمجموعه متفاوت است. سپس عملیات پنج بار تکرار می‌گردد. هر زمان، چهار زیرمجموعه از آن پنج زیرمجموعه، به عنوان داده‌های آزمایشی شناخته می‌شوند. اگر P، مجموعه تمام محصولات باشد، Li، مجموعه شامل محصولات کاربر ui است که دارای مجموعه آموزشی است و Ti، مجموعه‌ای شامل محصولات کاربر ui است. DOA به صورت زیر تعریف می‌گردد:



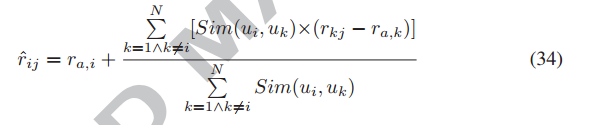
که در آن، U مجموعه تمام کاربران و |U|، سایز U است و DOAi، اندازه‌گیری برای کاربر ui با رتبه تعیین شده است. DOAi، به صورت زیر محاسبه می‌گردد.



به طوری که در آن، PPik و PPik نشانگر اولویت پیشنهادی برای کاربر ui و محصول Pj است. سپس DOAi به صورت زیر فرموله بندی می‌گردد:



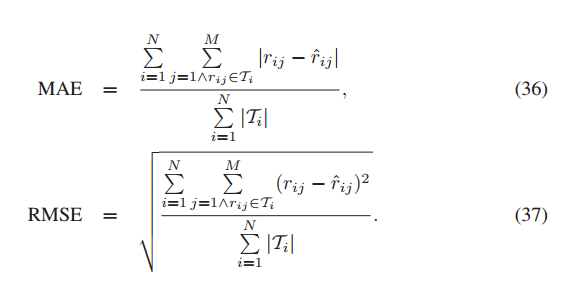
که در آن مجموعه محصولات کاربر ui، دارای مجموعه آموزشی تست است. توجه داشته باشید که مقدار بزرگ‌تر DOA، نشانگر اولویت سیستم پیشنهادی است. یک سیستم پیشنهادی دارای DOA برابر با 100% است. برای محاسبه MAE یا RMSE، ما اولویت پیشگویی را به امتیازات پیشگویی تبدیل می‌کنیم. اجازه دهید، rij، نشانگر امتیاز پیشگویی برای اولویت محصول pi برای کاربر uj باشد. سپس rij، به صورت زیر محاسبه می‌گردد.



طوری که ra,i، میانگین موجود در Li است و میانگین در Lk نیز وجود دارد که با rkj نشان داده شده است.



مجموعه‌های MAE و RMSE به صورت زیر تعیین می‌گردند:

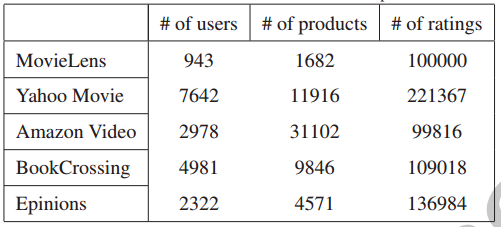


یک مقدار MAE و RMSE نشانگر اولویت یک سیستم توصیه گر است. یک سیستم توصیه گر ایده آل، دارای هیچ خطایی نیست و در آن MAE=0 است و RMSE=0 است. توجه داشته باشید که در هر یک از آزمایش‌ها زیر، لیست اولویت پیشنهادی از مجموعه آموزشی مشتق شده است. همچنین، هیچ overlap ای بین داده‌های آموزشی و تست نیست.

* 1. **مجموعه داده**

پنج مجموعه داده، MovieLens، yahoo Movie، Amazon Video، BookCrossing و Epinions برای آزمایش‌ها این بخش استفاده شده‌اند. مجموعه داده MovieLens [51]، به صورت عمومی موجود است و توسط موسسه تحقیقاتی GroupLens برای اجرا در سیستم های توصیه گر ارائه شده است. آن شامل 943 کاربر، 1682 محصول و 100000 رتبه دریافتی از کاربر است. به عبارت دیگر، N=943 و M=1682 است. هر سطر از مجموعه داده به صورت یک، سه تایی (ui,pj,rij) نشان داده شده است. بنابراین، این مجموعه داده دارای 100000 عنصر است. مجموعه داده Yahoo Movie، شامل اولویت جامعه Yahoo Movie است و شامل 7642 کاربر و 11916 محصول و 221367 اولویت برای هر کاربر است. هر سطر از آن، به صورت یک سه تایی (ui,pj,rij) نشان داده شده است. مجموعه داده ویدوئی آمازون، شامل تقسیم‌بندی ویدئوهای آمازون است که در طی 18 سال، شامل 143.7 میلیون تا سال 2014 است. آن شامل 2978 کاربر، 31102 محصول و 99816 اولویت برای هر کاربر است. هر سطر از آن نشانگر یک سه تایی است که در طی چهار هفته جمع‌آوری شده است. در نهایت مجموعه داده پایانی، شامل 4981 کاربر، 9846 محصول (کتاب) و 109018 اولویت برای هر کاربر است. هر عنصر در این مجموعه داده با یک سه تایی (Ui, Pj rij ) نمایش داده می‌شود که در آن . مجموعه داده Epinionos از Paolo massa در طی 5 هفته جمع‌آوری شده است که شامل 2322 کاربر، 4571 محصول و 136984 اولویت برای کاربر است. هر ورودی از مجموعه داده به عنوان یک سه تایی (ui,pj,rij) است که در آن rij عضو {1,2,3,4,5} است. ویژگی‌های این پنج مجموعه داده در جدول 5 خلاصه شده است.

جدول 5. ویژگی‌های مجموعه داده استفاده شده برای آزمایش‌ها

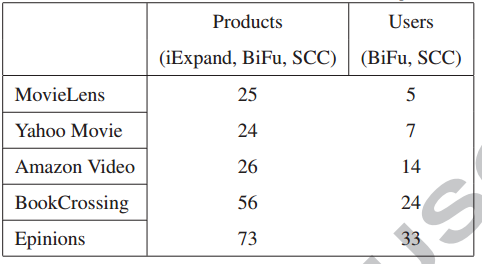


**6.2 نتایج و بحث**

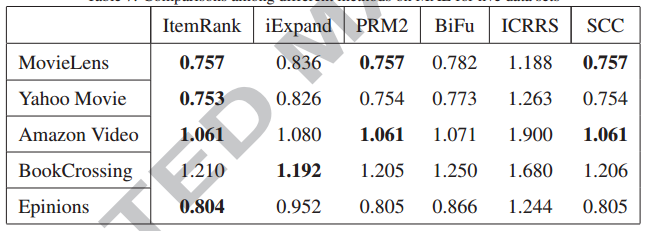
ما اثربخشی، دقت و کارایی، راه‌کار پیشنهادی SCC را توسط مقایسه با سایر روش‌های فیلترگذاری شامل Itemrank نشان داده‌ایم. همچنین iExpand [30]، PRM2 [33]، BiFu [34] و ICRRS [36]. همان طور که قبلاً توضیح داده شد، ItemRank از هیچ تکنیکی استفاده نمی‌کند و iExpand از روش LDA برای گروه‌بندی محصولات استفاده می‌کند و PRM2 از روش SVD برای تولید یک نمایش کاهش بعد یافته استفاده می‌کند و BiFu از روش K-means، برای خوشه‌بندی کاربران و محصولات بهره می‌گیرد و ICRRS از روش کاهشی که برای تجزیه ارزیابی استفاده می‌شود، بهره می‌گیرد. برای یک مقایسه امن، ما یک برنامه برای ItemRank، iExpand، PRM2، BiFu، ICRRS و SCC نوشتیم. تمام برنامه‌های نوشته شده از C++ در ویژوال استودیو 2013 بهره بردند. ما از کامپیوتری با Intel®Core(TW) i7-4790k CPU، با 4 GHz و 32GB RAM استفاده کردیم که دارای ویندوز7 برای اجرای برنامه‌ها بود. برای سهولت، ما از تعداد خوشه یکسانی برای گروه‌بندی کاربران و محصولات SCC، iExpand و BiFu استفاده کردیم. چون LDA و روش K-means نیاز به خوشه‌هایی برای تصمیم‌گیری پیشرفته دارند. در هر مورد ما از SCC برای یافتن تعداد خوشه‌ها استفاده کردیم. جدول 6 نشانگر تعداد خوشه‌ها برای خوشه‌بندی کاربران و محصولات تقسیم‌بندی شده درون SCC، iExpand و BiFu است. برای مثال، محصولات MovieLens دارای 25 خوشه برای iExpand، Bifu و SCC هستند.

جدول 7 و جدول 8 و جدول 9 نشانگر مقایسه MAE، RMSE و DOA هستند. بین ItemRankهای ایجادشده، Item Rank، iExpand و BoFu و ICRRS و SCC وجود دارند. توجه داشته باشید که این جداول دارای مقداری هستند که به صورت عملی جمع‌آوری شده است.

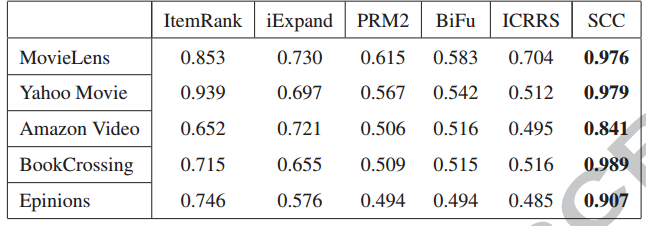
جدول 6. تعدادی از خوشه‌ها برای درگیر کردن کاربران و محصولات



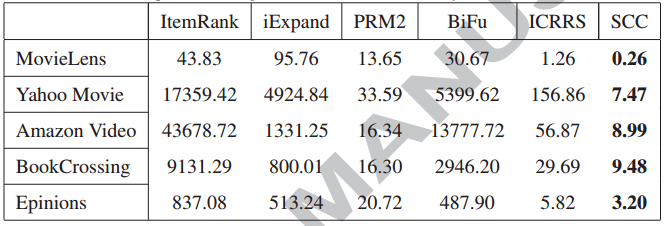
جدول 7. مقایسه بین روش‌های متفاوت MAE برای پنج مجموعه داده



جدول 8. مقایسه بین روش‌های متفاوت DOA برای پنج مجموعه داده

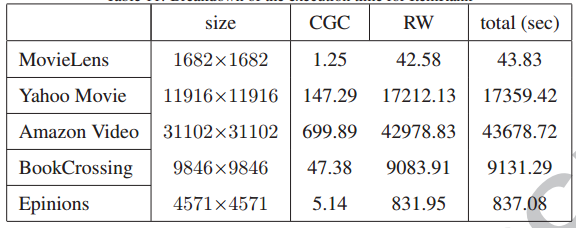


جدول 9. مقایسه بین روش‌های متفاوت کارایی (ثانیه) برای پنج مجموعه داده



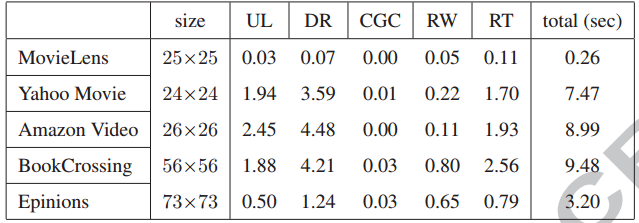
برای MAE و RMSE، مقادیر کوچک‌تر دارای کارایی بهتری هستند. در تضاد با آن، برای DOA، مقادیر بزرگ‌تر دارای اولویت بهتری هستند. همان طور که می‌توان دید، ItemRank و SCC دارای اجرای بهتری هستند و MAE و RMSE دارای مجموعه داده بهتری هستند. برای مثال، هر دوی ItemRank و SCC، دارای مقدار کمتری از 1.061 در MAE برای مجموعه داده آمازون هستند و هر دو دارای مقدار کمینه 1.044 در RMSE برای مجموعه داده Epinions هستند. PRM2، دارای اجرای بهتری از MAE و RMSE است. بنابراین برای DOA و SCC، اجرای بهتری در هر پنج شرایط مجموعه داده رخ می‌دهد. جدول 10 نشانگر مقایسه زمان اجرا، بین روش‌های متفاوت هستند. ما می‌توانیم ببنیم که SCC سریع تر از سایر روش‌ها اجرا می‌گردد. توجه داشته باشید که ItemRank، دارای کاهش بعد بهتری در مقایسه با سایرین است که در رابطه (4) نشان داده شده است. برای مثال، سایز 31102\*31102 برای مجموعه داده ویدئویی آمازون است. مصرف ماتریس بهینه برای آن توصیه می‌گردد. در نتیجه، ItemRank دارای زمان بیشتری در حدود 43678.72 ثانیه است که در جدول نشان داده شده است. به جای آن، SCC حالت خوشه‌بندی خودساخته را برای کاهش ابعاد استفاده کرده است و دارای 31102 گروه و 26 نوع محصول است. بنابراین، سایز ماتریس همبستگی استفاده شده در رابطه (26) نشان داده شده است. در نتیجه، SCC می‌تواند خیلی سریع تر اجرا گردد که دو حدود 8.99 ثانیه خواهد بود. BiFu و iExpand دارای کاهش بعد توسط K-means و LDA هستند که دارای مصرف زمان است. بنابراین، iExpand و BiFu کندتر از SCC اجرا می‌شوند. برای مثال، iExpand دارای 1331.25 ثانیه است. ماتریس همبستگی W، در رابطه (4)، دارای سایز M\*M در رابطه (26) است که دارای سایز q در q است. بنابراین SCC باید در مدت زمان (M/q)2 اجرا گردد که سریع‌تر از ItemRank است.

جدول 10. شکست زمان اجرا برای ItemRank



SCC باید در مدت زمانی برابر با 4500 برابر سریع تر از ItemRank برای MovieLens اجرا گردد. به هر حال، از جدول 10 دیده می‌شود که SCC، فقط 244 برابر سریع تر از ItemRank اجرا می‌گردد. یک دلیل این است که سایر گام‌ها دارای زمان بیشتری هستند. جدول 11 و جدول 12، نشانگر شکست در زمان اجرای ItemRank و SCC است. توجه داشته باشید که سایز ستون‌ها در جدول نشانگر سایز ماتریس مرتبط با آن است. ItemRank دارای دو گام است: ایجاد گراف همبستگی (CGC) و RandomWalk (RW) و SCC دارای پنج گام است: برچسب‌گذاری کاربر (UL) کاهش ابعاد (DR)، ایجاد گراف همبستگی (CGC)، Random Walk (RW) و re-transformation (RT). بررسی MovieLens نشان می‌دهد. در حالت کلی 43.83 ثانیه است و ItemRank دارای 1.25 در گام CGC است.

جدول 11. شکست زمان اجرای SCC



در طی کل 0.26 ثانیه، SCC در حدود 0.03 ثانیه است و در گام UL، 0.07 ثانیه می‌کشد. اگر ما به ستون RW نگاه کنیم، در این جدول این ستون‌ها در حدود 850 ثانیه زمان می‌خواهد.

انتخاب یک مقدار مناسب برای مقدار "رو" در SCC وابسته به فرآیند آزمون و خطا است. همان طور که در بخش 4 ذکر شده است، ممکن است در نتیجه الگوریتم خوشه‌بندی تأثیر زیادی در فرآیند توصیه گری داشته باشد. مقدار "رو" بزرگ‌تر نشان‌دهنده، تعداد خوشه کوچک‌تر و مقدار "رو" کوچک‌تر نشان‌دهنده تعداد خوشه بزرگ است.

**7. نتیجه‌گیری**

در یک سیستم فیلترینگ با همبستگی مشترک، مانند ItemRank، محصولات به عنوان یک روش مناسب شناخته شده است، اما معایبی بر آن وارد است. با این حال، معمولاً محصولات تجاری بسیار زیادی وجود دارند که قبلاً مورد بررسی قرار گرفته‌اند. ما یک روش خوشه‌بندی خودساخته برای کاهش ابعاد مرتبط با تعداد محصولات ارائه کرده‌ایم. در نتیجه، زمان پردازش اندازه‌گیری شده توسط روش ما برای تصمیم‌گیری در مورد یک روش توصیه خیلی کمتر شده است. نتایج آزمایش‌ها نشان می‌دهد که کارایی سیستم پیشنهادی بدون داشتن تأثیر بر روی کیفیت توصیه بهبود یافته است.

