ISSN: 2745-3673
http://www.Rcsj.ir

مجلـه‌ی علمی پژوهش در علوم رایانه (سال چهارم)
شماره ۱۳ / بهار ۱۳۹۸ / ص ۱۴- ۱

NWS_RS: شخصی‌سازی پیشنهادات براساس اطلاعات کاربران و معیار شباهت وزن‌دار جدید

مصطفی خلجی
دانشکده مهندسی کامپیوتر، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران.

نام نویسنده مسئول:
مصطفی خلجی

چکیده

با توجه به افزایش اطلاعات در فضای اینترنت، سیستم‌های توصیه‌گر به ابزاری توانمند برای پیداکردن سلایق، راهنمایی و هدایت کاربران به سمت اقلام مورد نیاز، تبدیل شده‌اند. پالایش همکارانه یکی از رویکردهای اصلی این سیستم‌ها در امر پیداکردن کاربران هم سلیقه و پیش‌بینی میزان علائق کاربران به اقلام خاص است. با افزایش تعداد کاربران و اقلام در سیستم، این رویکرد دچار مشکل مقیاس‌پذیری می‌شود. از طرفی با توجه به تنک بودن ماتریس امتیازات، عملکرد سیستم کاهش می‌یابد. از این رو، هدف این مقاله ارائه یک سیستم توصیه‌گر برای شخصی‌سازی پیشنهادات براساس اطلاعات کاربران به همراه معیارشباهت وزن‌دار جدید با نام NWS است. با بکارگیری اطلاعات جمعیتی کاربران از قبیل بازه‌ی سنی، می‌توان مشکل مقیاس‌پذیری را مدیریت کرد و پیشنهادات را برای کاربران شخصی‌سازی کرد. سیستم توصیه‌گر پیشنهادی مبتنی بر پالایش همکارانه کاربر محور بوده و با استفاده از درجه اطمینان تشابه کاربران به عنوان یک وزن در روند پیش‌بینی، عملکرد سیستم را ارتقا می‌دهد. نتایج آزمایشات سیستم توصیه‌گر پیشنهادی برروی مجموعه داده MovieLens صورت گرفته و ارزیابی سیستم با استفاده از معیارهای MAE، Accuracy، Precision، Recall و F1، بیانگر بهبود خطای پیش‌بینی، افزایش عملکرد و کارایی سیستم توصیه‌گر پیشنهادی نسبت به سایر روش‌های پالایش همکارانه‌ای است که از معیارهای شباهت مختلفی در امر پیشنهاددهی، استفاده می‌کنند. حداکثر نرخ خطای بهبود یافته سیستم ۱۷,۴ درصد و ۲۰,۲ درصد به ترتیب برای کاربران ۲۰ تا ۳۹ سال و کاربران ۴۰ تا ۶۰ سال می‌باشد.

واژگان کلیدی: سیستم توصیه‌گر، پالایش همکارانه، معیارشباهت وزن‌دار جدید *NWS*، شخصی‌سازی پیشنهادات.

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

با افزایش اطلاعات در اینترنت، فضای مجازی و خریدهای اینترنتی و تعاملات کاربران با یکدیگر، سیستم‌های توصیه‌گر[1](RS) جهت هدایت کاربران به سمت سلایق یا نیازهایی که دارند، در بیست سال اخیر و به ویژه در دهه‌ی اول قرن بیست و یکم مورد مطالعه قرار گرفته‌اند و پژوهش‌های بسیاری در این زمینه انجام شده است [۱]. این سیستم‌ها با توانایی کشف میزان ترجیحات کاربران و پیش‌بینی اولویت‌های آن‌ها، اقلامی را که احتمال می‌رود مورد پسند کاربر باشد را از بین حجم عظیمی از داده‌ها، پالایش کرده و با توصیه آن‌ها، از اتلاف وقت او جلوگیری می‌کند. امروزه سیستم‌های توصیه‌گر در کاربردهای متفاوتی از قبیل تجارت الکترونیک، توصیه موسیقی، فیلم، مقاله و غیره مورد استفاده قرار می‌گیرند [۲].

پالایش همکارانه[2](CF) یکی از مهم ترین رویکردهای بکاررفته در این سیستم‌ها می‌باشد. عملکرد این روش به این طریق است که براساس شباهت بین کاربران یا اقلام، پیش‌بینی و پیشنهاددهی را برای کاربران فعال[3] انجام می‌دهد. این الگوریتم با استفاده از ماتریس امتیازات کاربر-قلم، کاربران یا اقلام مشابه را برای پیش‌بینی میزان ترجیحات کاربران فعال نسبت به اقلامی که خریداری یا مشاهده نکرده‌اند، پیدا می‌کند و در آخر لیستی از پیشنهادات را به آنها توصیه می‌کند. ماتریس امتیازات کاربر-قلم، متشکل از امتیازات کاربران به هریک از اقلام است که به طور معمول برخی از کاربران تمایلی به امتیاز دادن به اقلامی که مشاهده یا خریداری کرده‌اند، ندارند. این رویکرد از مشکلاتی همچون تنکی‌داده[4] و مقیاس‌پذیری[5] رنج می‌برند. هنگامی که سیستم توصیه‌گر با کمبود امتیازات از سوی سوابق کاربر مواجه می‌شود، مشکل تنکی‌داده بوجود می‌آید که تاثیر بسزایی بر افزایش خطای پیش‌بینی و کاهش عملکرد سیستم دارد. از طرفی دیگر با افزایش تعداد کاربران و اقلام، همچنین محاسبه معیار شباهت برای تک‌تک کاربران فعال با کاربران دیگر، سیستم دچار مشکل مقیاس پذیری می‌شود [۳]. برای پیدا کردن تشابه کاربران نسبت به کاربران فعال، معیارهای شباهت مختلفی استفاده می‌شوند که عبارتند از: Pearson، Jaccard، SPC، CPC، MSD [۴-۸]. از این رو هدف این مقاله، طراحی یک سیستم توصیه‌گر مبتنی بر رویکرد پالایش همکارنه، برای شخصی‌سازی پیشنهادات براساس اطلاعات جمعیتی کاربران(سن) به همراه معیار شباهت وزن‌دار جدید (NWS)[6] است که بتوان خطای پیش‌بینی و عملکرد سیستم را بهبود داد. باتوجه به افزایش روز افزون کاربران و اقلام در فضای وب، می‌توان با تفکیک کردن کاربران براساس بازه‌سنی، مساله مقیاس‌پذیری را مدیریت و پیشنهاددهی را برای هر کاربر شخصی‌سازی کرد. هنگامی که تعداد اقلام مشترک بین کاربران فعال و همسایه زیاد باشد، کاربر همسایه‌ای با بیشترین اقلام مشترک از نظر سلیقه احتمال دارد بسیار مورد اطمینان باشد. این امر را می‌توان به عنوان یک ضریب وزنی در سیستم پیشنهاددهی استفاده کرد که منجر به بهبود خطای پیش‌بینی، افزایش کارایی و عملکرد سیستم توصیه‌گر پیشنهادی می‌شود.

ساختار مقاله به این صورت می‌باشد که در بخش ۲ به مروری برکارهای محققین پرداخته می‌شود و در بخش ۳ سیستم توصیه‌گر پیشنهادی معرفی می‌شود. بخش ۴، بخش ارزیابی سیستم توصیه‌گر پیشنهادی می‌باشد که در آن به نتایج آزمایشات و مقایسه با روش‌های دیگر پرداخته می‌شود. در نهایت در بخش آخر نیز نتیجه‌گیری ارائه می‌گردد.

۱- مروری برکارهای محققین

سیستم‌های توصیه‌گر در ابتدا توسط گلدبرگ و همکارانش معرفی شدند [۹]. سیستم‌های توصیه‌گر ابزاری برای هرچه توانمند کردن کاربران در پیدا کردن اقلام خاص، در فضای مجازی هستند. این سیستم‌ها با پیدا کردن سلایق کاربران پیشنهاداتی را برایشان مهیا می‌سازند. پالایش همکارانه یکی از پرکاربردترین رویکردهای بکارگرفته شده در پیدا کردن سلایق کاربران است. این رویکرد با استفاده از معیار شباهت‌های مختلف، کاربرانی که از نظر سلیقه به کاربرفعال نزدیک هستند را پیدا و براساس آن میزان علاقه کاربر فعال به اقلام خاص را پیش‌بینی می‌کند. اگرچه این روش از مشکلاتی از قبیل شروع سرد، تنکی‌داده و مقیاس پذیری رنج می‌برد اما فهم و پیاده سازی آنها بسیار راحت است و از مدل‌های پایه در سیستم‌های توصیه‌گر می‌باشند. از این رو روش‌های جدیدی مبتنی بر پالایش همکارانه توسط محققین برای بهبود عملکرد سیستم‌های توصیه‌گر ارائه گردیده است [۱۰].

پالایش همکارانه به نوبه خود دارای دو روش اصلی مدل محور[7] و حافظه محور[8] است. رویکرد مدل محور با استفاده از روش‌های هوشمند یادگیری ماشین، رفتار کاربران را نسبت به سوابق خودشان مدل سازی می‌کند. از طرفی دیگر روش حافظه محور، با استفاده از

1. Recommender Systems
2. Collaborative Filtering
3. Active Users
4. Data Sparsity
5. Scalability
6. New Weighted Similarity
7. Model-Based

ISSN: 2745-3673
http://www.Rcsj.ir

رویکرد نزدیک همسایه، به دنبال پیدا کردن کاربرانی است که از نظر سلیقه و ترجیحات به کاربر فعال نزدیک باشند [۱]. امروزه بسیاری از محققین با ترکیب این دو روش به حل مشکلات مطرح شده می پردازند.

برای بهبود عملکرد سیستم توصیه‌گر، یک معیار شباهت جدید با نام PIP[۹] معرفی شد که توانست مشکل شروع سرد را حل کند. در این معیار از اختلاف بین امتیازات دو کاربر به همراه بررسی امتیازات هر دو کاربر در شرایط سازش(توافق[۱۰]) و عدم سازش(عدم توافق) با یکدیگر، استفاده شد. روش مذکور دارای ضریب جریمه بوده و در شرایطی که امتیازات کاربران در وضعیت عدم توافق با یکدیگر باشند، اعمال می‌شود [۱۱]. لیو و همکارانش یک معیار شباهت اکتشافی جدید با نام NHSM[۱۱] ارائه کردند که بجای اثر شدید[۱۲] روش [۱۱]، تکینگی[۱۳] امتیازات را هنگام انتخاب کاربران همسایه کاربر فعال، در نظر گرفتند [۱۲]. بلوگین و همکارانش روش‌هایی را برای بهبود عملکرد سیستم‌های توصیه‌گر معرفی کردند. آنها از روش‌های وزن‌دار شناخته شده با نام HW[۱۴] [۱۳] وMW[۱۵] [۱۴] با عدم بکارگیری از معیارهای شباهت، کاربرانی که از نظر سلیقه به کاربر فعال نزدیک بودند را پیدا و در روند پیش‌بینی سیستم پیشنهادی خود استفاده کردند. سیستم آنها از نظر پیچیدگی زمانی نسبت به سایر روش‌هایی که از معیارهای شباهت استفاده می‌کردند، عملکرد قابل قبولی را در پی داشت [۱۵]. چوی و همکارانش در سال ۲۰۱۳ یک معیار شباهت جدید برای انتخاب همسایگان برای هر قلم فعال در پالایش همکارانه ارائه کردند که امتیاز کاربر به قلم را براساس معیار خود وزن‌دار می‌کردند [۱۶]. جواری و همکارانش در سال ۲۰۱۴ یک سیستم توصیه‌گر مبتنی بر پالایش همکارانه و تخصیص منابع ارائه کردند. آنها با استفاده از روش تخصیص منابع توانستند درجه اعتماد هر کاربر را براساس میزان شباهت بدست آمده، محاسبه کنند و در نتیجه عملکرد سیستم را نسبت به سایر روش‌های معمول بهبود بخشند [۱۷]. ژانگ و همکارانش در سال ۲۰۱۶ یک روشی برای بهبود توانایی پیدا کردن کاربران مشابه و قابل اطمینان به کاربر فعال را ارائه دادند، هدف آنها ارائه یک سیستم توصیه‌گر موثر مبتنی بر مدل برای حل مشکل تنکی‌داده بود [۱۸].

پارک و همکارانش در سال ۲۰۱۵ یک الگوریتم پالایش همکارانه سریع با استفاده از نزدیک ترین گراف همسایه برای کاهش دادن مشکل پیچیدگی زمانی، ارائه کردند. روش آنها RCF[۱۶] نام دارد که فرآیند پردازش پیدا کردن نزدیک‌ترین همسایه را نسبت به پالایش همکارانه معمول، معکوس می‌کند [۱۹]. خلجی در سال ۱۳۹۶یک سیستم توصیه‌گر ترکیبی فیلم با استفاده از شبکه عصبی و تخصیص منابع ارائه کرد. او توانستد مشکل مقیاس‌پذیری را با استفاده از شبکه عصبی نگاشت خود سازمان‌ده و روش پیش‌بینی پیوند در شبکه‌های اجتماعی، حل نماید. روش او با جداسازی کاربران نسبت به اطلاعات جمعیتی[۱۷]شان و کشف سلایق هر یک از کاربران در اقلام و ژانرهای خاص، کارایی سیستم را بهبود بخشید [۲۰].

کوهی و همکارش در سال ۲۰۱۶، با بکارگیری خوشه‌بندی فازی و روش فازی‌زدایی Max، به صورت تخصیص کاربران به تمام خوشه‌ها با درجه عضویت متفاوت و استفاده از معیار شباهت Pearson برای پیداکردن نزدیک‌ترین همسایه، نشان دادند که عملکرد سیستم آنها نسبت به استفاده از روش‌های K-means و SOM[۱۸] بهبود بخشیده شده است [۲۱]. بلاسل و همکارانش در سال ۲۰۱۸ یک سیستم توصیه‌گر مقیاس‌پذیر مبتنی بر رویکرد پالایش همکارانه معرفی کردند. آنها با استفاده از الگوریتم خوشه‌بندی جداسازی-ادغام توانستند زمان و دقت سیستم پیشنهادی‌شان را بهبود دهند [۲۲]. کنت و همکارانش در سال ۲۰۱۸ یک روشی برای مشخص کردن انتخاب مرکز اولیه خوشه در عملیات خوشه‌بندی K-means معرفی کردند. روش آنها توانست مشکل تنکی‌داده را حل کردند [۲۳]. خلجی و دادخواه درسال ۱۳۹۷ یک سیستم توصیه‌گر ترکیبی با نام FNHSM_HRS معرفی کردند. آنها با بکارگیری روش خوشه‌بندی فازی و استفاده از یک معیار شباهت اکتشافی مشکل مقیاس پذیری را حل نمودند. سیستم آنها در ابتدا رفتار کاربران را براساس روش‌های فازی مدل کرد، سپس برای پیدا کردن نزدیک‌ترین همسایه از معیار شباهت اکتشافی استفاده کرد [۳]. خداوردی و همکارانش در سال ۱۳۹۷ یک سامانه توصیه‌گر ترکیبی فیلم مبتنی بر خوشه‌بندی و محبوبیت را ارائه کردند. سیستم آنها با استفاده از روش خوشه‌بندی K-means، کاربران مشابه را در خوشه‌های مجزا قرار داد و با بکارگیری روش محبوبیت امتیازات، میزان علاقه کاربر فعال را برای فیلم‌هایی که مشاهده نکرده بود، پیش‌بینی کرد و در آخر روش آنها توانست مشکل مقیاس پذیری و تنکی‌داده را حل کند [۱۰]. خلجی و محمّدنژاد در سال ۲۰۱۹ یک

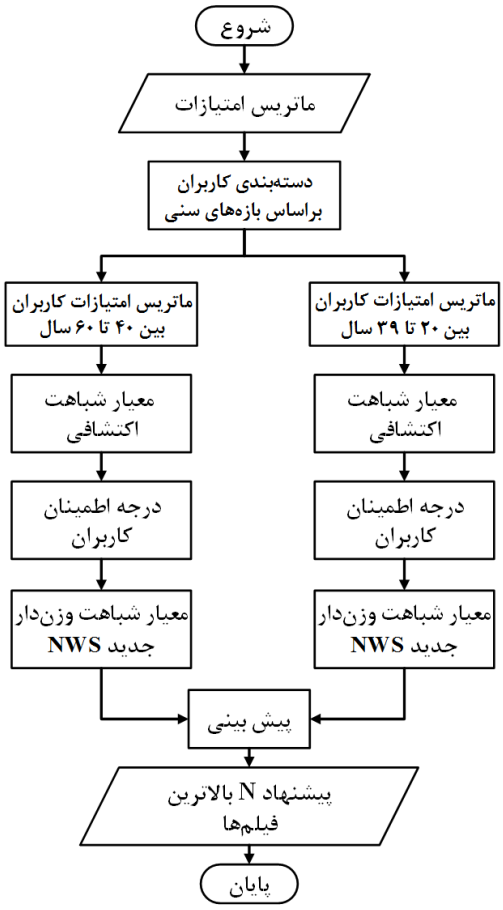
8. Memory-Based
9. Proximity, Impact, Popularity
10. Agreement
11. New Heuristic Similarity Measure
12. Impact
13. Singularity
14. Herlocker Weighting
15. McLaughlin Weighting
16. Reversed CF
17. Demographic Information
18. Self-Organizing Map

ISSN: 2745-3673
http://www.Rcsj.ir

سیستم توصیه‌گر ترکیبی فیلم با نام FCNHSMRA_HRS معرفی کردند. سیستم آنها با ترکیب روش‌های مدل محور و حافظه محور در پالایش همکارانه به همراه یکی از روش‌های پیش‌بینی پیوند، توانستند عملکرد سیستم خود را نسبت به روش قبلی خود(FNHSM_HRS) و سایر روش‌های مرسوم بهبود بخشند [۲۴]. وانگ و همکارانش در سال ۲۰۱۹ یک روشی جدید با نام CDIE[۱۹] معرفی کردند. آنها از روش Co-Clustering برای استخراج همبستگی اقلام و پالایش کردن اختشاش‌ها استفاده کردند. روش آنها توانست مشکل تنکی‌داده را رفع کند [۲۵].

۲- سیستم توصیه‌گر پیشنهادی

در شکل شماره ۱ ساختار سیستم توصیه‌گر پیشنهادی نشان داده شده است. این ساختار مبتنی بر الگوریتم پالایش همکارانه حافظه محور و روش کاربر محور است.



شکل شماره ۱: ساختار سیستم توصیه‌گر پیشنهادی

سیستم توصیه‌گر پیشنهادی دارای ماتریس امتیازات کاربر-فیلم است که بیانگر تعدادی امتیاز کاربران به فیلم‌ها است. در این مقاله مجموعه کاربران با $U = [u_1, u_2, \dots, u_M]$، فیلم‌ها با $I = [i_1, i_2, \dots, i_N]$ و ماتریس امتیازات با نام RMatrix معرفی می‌شوند. سایز RMatrix برابر است با تعداد کاربران × تعداد فیلم‌ها $(M * N)$.

۲-۱- دسته‌بندی کاربران براساس بازه‌های سنی

در این بخش تمامی کاربران براساس اطلاعات جمعیتی ایشان، دسته‌بندی می‌شوند. اطلاعات جمعیتی کاربران شامل چندین ویژگی از قبیل جنسیت، حرفه(شغل)، میزان تحصیلات، سن و کدپستی است. باتوجه به گذر زمان و افزایش سن، سلیقه افراد تغییرپذیر است. از

19. Cross-Domain Item Embedding

ISSN: 2745-3673
http://www.Rcsj.ir

این رو، کاربرانی که در بازه سنی جوان دسته‌بندی می‌شوند، از نظر سلیقه و نوع امتیازدهی با کاربرانی که در بازه سنی میانسال قرار دارند، متفاوت است. به طور مثال، کاربران جوان تمایل دارند، فیلم‌هایی با ژانر کنشی مشاهده کنند، اما کاربران میانسال ممکن است به این نوع از ژانر علاقه نشان ندهند و ژانر دیگری را برای خود برگزینند. در سیستم توصیه‌گر پیشنهادی کاربران فقط براساس بازه‌های سنی ۲۰ تا ۳۹ سال و ۴۰ تا ۶۰ سال از همدیگر تفکیک شدند. دسته‌بندی کاربران باعث تسریع در روند پیش‌بینی و پیشنهاددهی می‌شود. علاوه بر این کاربران فعال هنگام ورود به سیستم، تنها می‌توانند از نظرات کاربران هم سن خود برای فیلم‌هایی که مشاهده ننموده‌اند، توصیه‌هایی را دریافت کنند. در پایان این بخش، دو ماتریس امتیازات برای بازه‌های سنی ۲۰ تا ۳۹ و ۴۰ تا ۶۰ سال به نام‌های RMatrixA1 و RMatrixA2 با ابعاد $m * n$ از ماتریس RMatrix به وجود می‌آیند. RMatrixA1 $\cup$ RMatrixA2 = R'Matrix.

۲-۲- پیدا کردن K نزدیک‌ترین همسایه

سیستم توصیه‌گر پیشنهادی ماتریس‌های RMatrixA1 و RMatrixA2 را به عنوان ورودی دریافت کرده و براساس بازه‌های سنی، ماتریس متعلق به آن کاربرفعال را برای محاسبه K نزدیک‌ترین همسایه انتخاب می‌کند. در گام بعد، سیستم برای پیدا کردن کاربرانی که از نظر سلایق با کاربرفعال مشابه هستند را با استفاده از معیار شباهت اکتشافی [۱۲] طبق رابطه (۱) محاسبه می‌کند.

$$NHSM\_Sim(u,v) = JPSS\_Sim(u,v) \,.\, URP\_Sim(u,v) \quad (۱)$$

این معیار دارای دو ضریب اصلی است که هر یک از ضرایب ترکیبی از تعدادی معیارهای شباهت معمول است. از این رو برای محاسبه معیار شباهت اکتشافی مذکور، در ابتدا می‌بایست معیار شباهت $JPSS\_Sim(u,v)$ محاسبه شود. این معیار از دو معیار شباهت دیگر مشتق شده است که در روابط (۲) و (۳) ذکر شده است.

$$JPSS\_Sim(u,v) = PSS\_Sim(u,v) \,.\, Jaccard'\_Sim(u,v) \quad (۲)$$

$$Jaccard'\_Sim(u,v) = \frac{|I_u \cap I_v|}{|I_u| \times |I_v|} \quad (3)$$

$|I_u \cap I_v|$ بیانگر تعداد فیلم‌های مشترکی است که کاربر u و v مشاهده نموده‌اند و $|I_u|$ بیانگر تعداد فیلم‌هایی که کاربر فعال (u) امتیاز داده است و $|I_v|$ بیانگر تعداد فیلم‌هایی که کاربر همسایه (v) امتیاز داده است. معیار PSS_Sim از طریق رابطه (۴) بدست می‌آید.

$$PSS\_Sim(r_{u,i}, r_{v,i}) = Proximity(r_{u,i}, r_{v,i}).Significance(r_{u,i}, r_{v,i}).Singularity(r_{u,i}, r_{v,i}) \quad (۴)$$

$Proximity$، $Significance$ و $Singularity$ از طریق روابط (۵) و (۶) و (۷) بدست می‌آید.

$$Proximity(r_{u,i}, r_{v,i}) = 1 - \frac{1}{1 + \exp(-|r_{u,i} - r_{v,i}|)} \quad (۵)$$

$$Significance(r_{u,i}, r_{v,i}) \frac{1}{1 + \exp(-|r_{u,i} - r_{med}| \,.\, |r_{v,i} - r_{med}|)} \quad (۶)$$

$$Singularity(r_{u,i}, r_{v,i}) = 1 - \frac{1}{1 + \exp(-|\frac{r_{u,i} + r_{v,i}}{2} - \mu_i|)} \quad (۷)$$

$r_{u,i}$ امتیاز فیلم i توسط کاربرفعال u و $r_{v,i}$ امتیاز فیلم i توسط کاربر v است. r_{med} هم میانگین امتیازاتی است که کاربر می‌تواند برای امتیازدهی به یک فیلم خاص تخصیص دهد. ماتریس امتیازدهی در سیستم توصیه‌گر پیشنهادی دارای محدوده امتیازات بین اعداد ۱ تا ۵ است که متوسط آن عدد ۳ می‌باشد. μ_i هم میانگین امتیازات فیلم i توسط تمام کاربران است.

آخرین مرحله از رابطه (۱)، محاسبه معیار شباهت URP_Sim است که با استفاده از رابطه (۸) بدست می‌آید.

$$URP\_Sim(u,v) = 1 - \frac{1}{1 + \exp(-|\mu_u - \mu_v| \,.\, |\sigma_u - \sigma_v|)} \quad (۸)$$

ISSN: 2745-3673
http://www.Rcsj.ir

μ_u میانگین امتیازات و σ_u انحراف معیار کاربرفعال است که مقدار σ_u طبق رابطه (۹) بدست می‌آید.

$$\sigma_u = \sqrt{\sum_{i \epsilon I_u} \left(r_{u,i} - \overline{r_u}\right)^2 / |I_u|} \qquad (۹)$$

اکثرا معیارهای شباهت موجود به طور مستقیم از میزان ترجیحات(امتیازات) کاربران، برای محاسبه تشابه نزدیک‌ترین کاربران همسایه استفاده می‌کنند و در آخر قادر به پیش‌بینی درست تشابه بین کاربران نیستند. از این رو هنگامی که تعداد فیلم‌های مشترک بین کاربران کم باشد، نتایج تشابه بدست آمده کاربران، قابل اطمینان نیست. فرض کنید تعداد فیلم‌های مشترک بین دو کاربر u و v، ۵ و تشابه آنها نسبت به یکدیگر ۰,۸ باشد و در مقابل تعداد فیلم‌های مشترک بین دو کاربر u و v، ۱۰۰ و تشابه آنها ۰,۶ باشد، نمی‌توان به اینگونه معیارها برای پیدا کردن K نزدیکترین همسایه اطمینان کرد. بنابراین در این مقاله، فرض می‌شود کاربرانی که تعداد فیلم‌های مشترک زیادی با کاربر فعال دارند، نسبت به کاربرانی که تعداد فیلم‌های مشترک کمتری دارند، از نظر سلیقه به کاربرفعال نزدیک‌تر و قابل اطمینان‌تر هستند و می‌توان به عنوان یک وزن در معیارهای شباهت استفاده کرد. توابع مختلفی برای تبدیل درجه همپوشانی فیلم‌ها به یک وزن معرفی شده‌اند [۱۳] و [۱۴]. در این توابع از یک حدآستانه قابل اطمینان برای همپوشانی فیلم‌های مشترک بین کاربران استفاده می‌شود. برای محاسبه درجه اطمینان تشابه کاربران نسبت به یکدیگر از رابطه (۱۰) استفاده می‌شود.

$$\boldsymbol{Reliability\_Degree(u,v) = \frac{max(\beta, |I_u \cap I_v|)}{\beta}} \qquad \boldsymbol{(10)}$$

β حدآستانه میزان همپوشانی فیلم‌های مشترک بین دو کاربر است. عملکرد این تابع به این صورت است که کاربرانی که تعداد فیلم‌های مشترک بیشتر از حد آستانه داشته باشند دارای اهمیت و وزن بیشتری در روند پیش‌بینی میزان ترجیحات کاربران فعال، هستند. در گام آخر این بخش، با ترکیب معیار شباهت اکتشافی به همراه تابع وزن‌دار(درجه اطمینان)، یک معیار شباهت وزن‌دار جدید به نام NWS طبق رابطه (۱۱) بدست می‌آید.

$$\boldsymbol{NWS(u,v) = NHSM\_Sim(u,v).Reliability\_Degree(u,v)} \qquad \boldsymbol{(11)}$$

پس از محاسبه معیار شباهت وزن‌دار جدید NWS، دو ماتریس جداگانه به نام‌های NWS_A1 و NWS_A2 در ابعاد $m * n$ به ترتیب برای کاربران در بازه‌های سنی ۲۰ تا ۳۹ و ۴۰ تا ۶۰ سال ایجاد می‌شوند. این ماتریس‌ها به صورت متقارن هستند و در هر درایه میزان تشابه وزن‌دار کاربران به کاربر فعال را مشخص می‌کنند. به طور مثال اگر u_1 کاربر فعال، u_2 کاربر همسایه و شباهت وزن‌دار بین کاربرu_1 و u_2، ۰,۳۶۵ باشد، باتوجه متقارن بودن ماتریس‌ها، شباهت بین کاربر u_2 با کاربر فعال u_1 همان ۰,۳۶۵ است. میزان تشابه بدست آمده $NWS(u,v)$ در بازه (۰و۱) است، چون خروجی هریک از روابط فوق در بازه ۰ تا ۱ است.

۲-۳- پیش‌بینی

در این مرحله تعداد K نزدیک‌ترین کاربران همسایه براساس بالاترین درجه شباهت وزن‌دار به کاربر فعال در بازه‌های سنی مربوطه، مشخص و انتخاب می‌شوند. مقدار K انتخابی برای همسایه‌های کاربرفعال، ۲۰۰ در نظر گرفته شده است. از این رو با توجه به رابطه (۱۲) عملیات پیش‌بینی امتیازات فیلم‌های مشاهده نشده براساس پالایش همکارانه کاربرمحور در سیستم توصیه‌گر پیشنهادی، محاسبه می‌شود.

$$\boldsymbol{Predict(u,i) = \mu_u + \frac{1}{\sum_{j=1}^{K} |NWS(u,v_j)|} \cdot \sum_{j=1}^{K} \left(r_{v_j,i} - \mu_v\right) . NWS(u,v_j)} \qquad \boldsymbol{(12)}$$

به طوری که u کاربر فعال، v_j کاربر همسایه و i فیلمی می‌باشد که سیستم توصیه‌گر پیشنهادی قرار است میزان ترجیح(امتیاز) کاربرفعال را برای آن پیش‌بینی کند، همچنین K تعداد کل کاربران همسایه، $r_{v_j,i}$ امتیاز کاربر v_j به فیلم i و μ_u میانگین امتیازات کاربر فعال است.

ISSN: 2745-3673
http://www.Rcsj.ir

۲-۴- پیشنهاددهی

در این بخش لیستی از پیشنهادات براساس روش Top-N برای کاربران فعال ایجاد می‌شود. مقدار N در این مقاله به ترتیب ۵، ۱۰، ۱۵، ۲۰ و ۳۰ در نظر گرفته شده است.

۳- ارزیابی سیستم توصیه‌گر پیشنهادی

داده‌های ورودی سیستم توصیه‌گر پیشنهادی مجموعه داده فیلم MovieLens شامل ۹۴۳ کاربر و ۱۶۸۲ فیلم با ۱۰۰ هزار امتیاز کاربران به فیلم‌ها، در نظر گرفته شده است [۲۴]. بازه امتیازدهی در این مجموعه داده از اعداد ۱ تا ۵ است که به ترتیب ۱ نشان دهنده مورد پسند نبودن و عدد ۵ نشان دهنده مورد پسند بودن کاربر به یک فیلم خاص می‌باشد. تعداد کاربران در بازه سنی ۲۰ تا ۳۹ سال و ۴۰ تا ۶۰ سال در ماتریس امتیازات RMatrixA1 و RMatrixA2 برای ارزیابی به ترتیب ۵۷۷ و ۲۷۱ کاربر می‌باشد.

برای ارزیابی عملکرد سیستم، از روش 5-fold Cross-Validation، استفاده شده است که ۸۰ درصد داده‌ها برای آموزش سیستم توصیه‌گر پیشنهادی و ۲۰ درصد داده‌ها برای آزمایش سیستم بکارگرفته می‌شود. روش مذکور شامل ۵ مرحله مستقل از هم می‌باشد که در هر مرحله داده‌های آموزش و آزمایش هر Fold به سیستم تزریق می‌شود [۲۵]. ارزیابی سیستم براساس معیارهای میانگین خطای مطلق[۲۰] (MAE)، دقت[۲۱]، صحت[۲۲] و بازیابی[۲۳] طبق جدول (۱) و روابط (۱۳-۱۷) برروی داده‌های آزمایش محاسبه شده است [۲۶]. مقدار حد آستانه برای محاسبه معیارهای ارزیابی ۴ در نظر گرفته شد، مقدار امتیاز ۴ تا ۵ بیانگر مورد پسند بودن(دوست داشتن) و مقدار ۱ تا ۴ بیانگر مورد پسند نبودن(دوست نداشتن) کاربرفعال برای فیلم است. علاوه براین مقدار حد آستانه $\boldsymbol{\beta}$ برای محاسبه درجه اطمینان تشابه کاربران در سیستم توصیه‌گر پیشنهادی ۵۰ در نظر گرفته شده است. این سیستم، با ۵ روش معمول و پرکاربرد در این حوزه همچون Pearson، Jaccard، SPCC، CPC و MSD مورد مقایسه و ارزیابی قرار گرفت، نتایج آزمایشات در شکل‌های (۲) تا (۱۱) بیانگر بهبود خطای پیش‌بینی و عملکرد سیستم نسبت به سایر معیارهای شباهت در هر دو بازه سنی است.

جدول ۱: ماتریس اغتشاش

Actual / Predicted	Negative	Positive
Negative	A	B
Positive	C	D

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{r}_{u,i} - r_{u,i}|} \quad (۱۳)$$

$$\boldsymbol{Accuracy = \frac{A + D}{A + B + C + D}} \quad (۱۴)$$

$$\boldsymbol{Precision = \frac{D}{B + D}} \quad (۱۵)$$

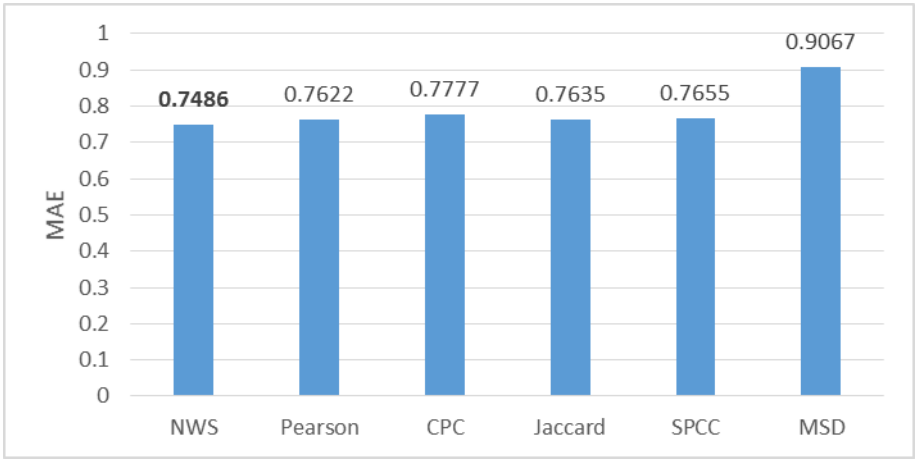
$$\boldsymbol{Recall = \frac{D}{C + D}} \quad (۱۶)$$

$$\boldsymbol{F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}} \quad (۱۷)$$

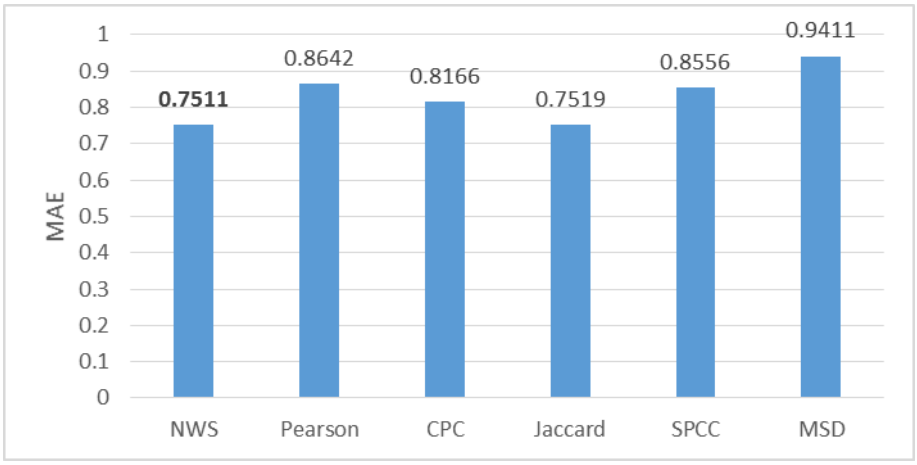
در شکل (۲) نرخ خطای سیستم توصیه‌گر پیشنهادی برای کاربران بین ۲۰ تا ۳۹ سال نسبت به سایر معیارهای شباهت به ترتیب ۱,۸درصد، ۳,۷درصد، ۱,۹۵درصد، ۲,۲۱درصد و ۱۷,۴درصد بهبود یافته و در شکل (۳) برای کاربران بین ۴۰ تا ۶۰ سال به ترتیب ۱۳,۱درصد، ۸ درصد، ۰,۱۱درصد، ۱۲,۲درصد و ۲۰,۲درصد بهبود یافته است. دیگر نتایج آزمایشات در شکل‌های (۴) تا (۱۱)، در بیشتر معیارهای ارزیابی و شرایط، بیانگر بهبود عملکرد و کارایی سیستم توصیه‌گر پیشنهادی است.

20. Mean Absolute Error
21. Accuracy
22. Precision
23. Recall

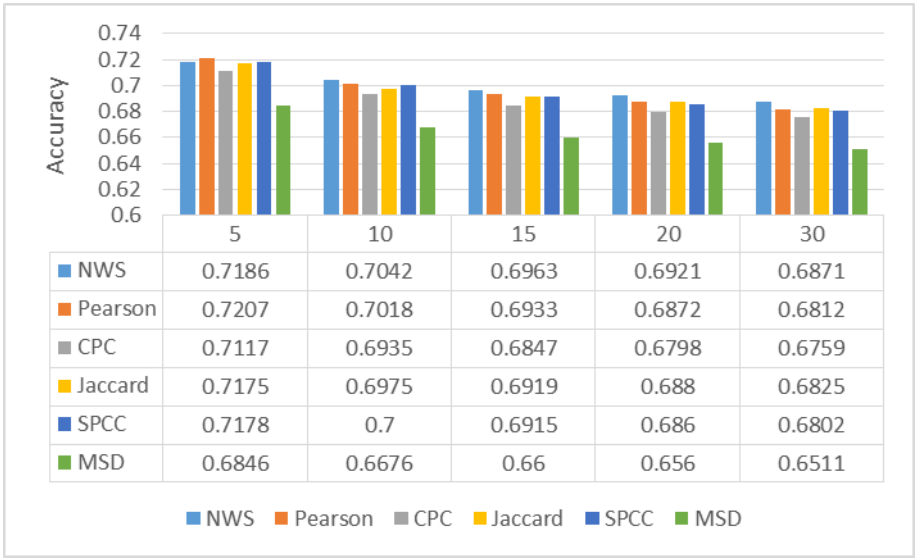
ISSN: 2745-3673
http://www.Rcsj.ir



شکل ۲: میانگین مطلق خطای پیش‌بینی سیستم برای بازه‌سنی ۲۰ تا ۳۹ سال



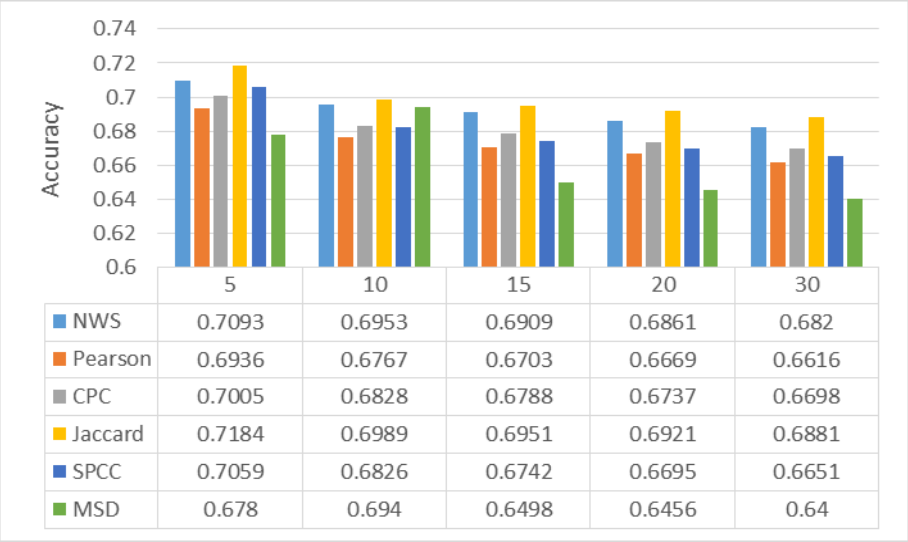
شکل ۳: میانگین مطلق خطای پیش‌بینی سیستم برای بازه‌سنی ۴۰ تا ۶۰ سال



	5	10	15	20	30
NWS	0.7186	0.7042	0.6963	0.6921	0.6871
Pearson	0.7207	0.7018	0.6933	0.6872	0.6812
CPC	0.7117	0.6935	0.6847	0.6798	0.6759
Jaccard	0.7175	0.6975	0.6919	0.688	0.6825
SPCC	0.7178	0.7	0.6915	0.686	0.6802
MSD	0.6846	0.6676	0.66	0.656	0.6511

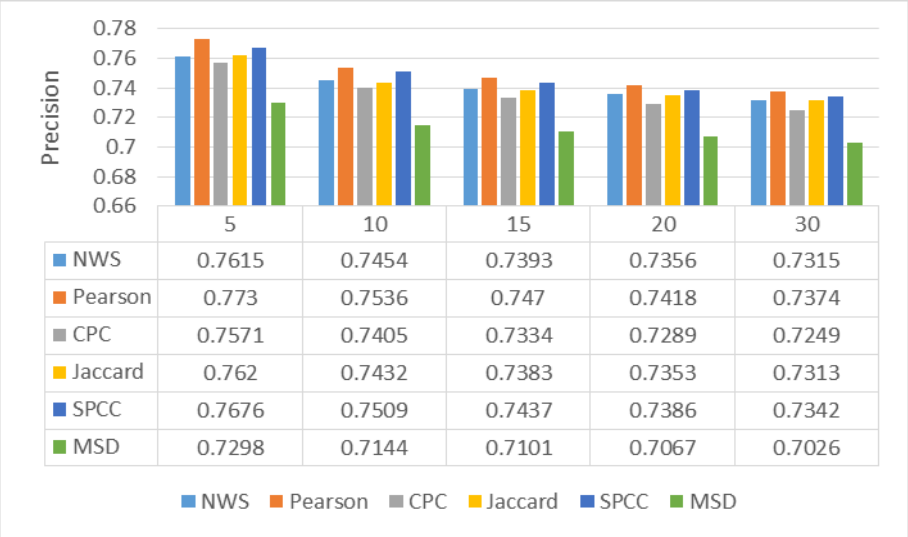
شکل ۴: دقت سیستم برای بازه‌سنی ۲۰ تا ۳۹ سال

ISSN: 2745-3673
http://www.Rcsj.ir



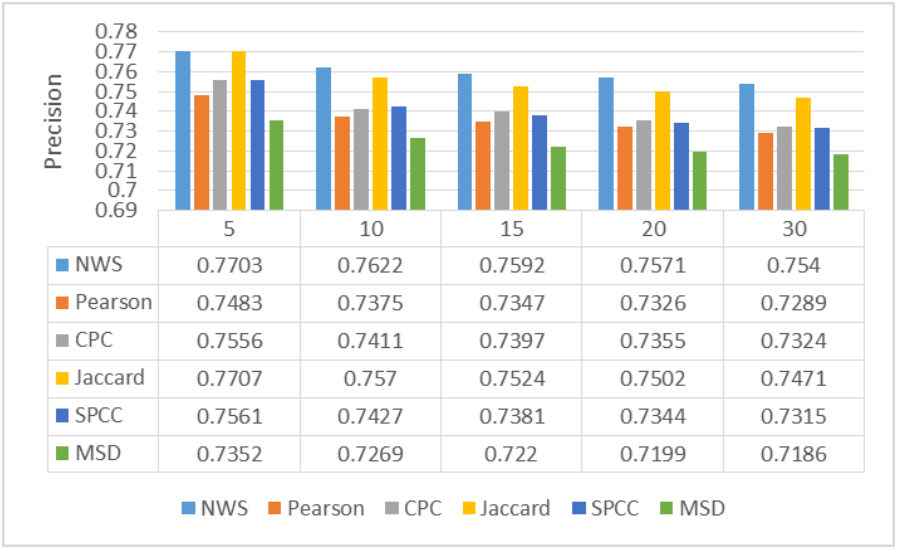
	5	10	15	20	30
NWS	0.7093	0.6953	0.6909	0.6861	0.682
Pearson	0.6936	0.6767	0.6703	0.6669	0.6616
CPC	0.7005	0.6828	0.6788	0.6737	0.6698
Jaccard	0.7184	0.6989	0.6951	0.6921	0.6881
SPCC	0.7059	0.6826	0.6742	0.6695	0.6651
MSD	0.678	0.694	0.6498	0.6456	0.64

شکل ۵: دقت سیستم برای بازه‌سنی ۴۰ تا ۶۰ سال



	5	10	15	20	30
NWS	0.7615	0.7454	0.7393	0.7356	0.7315
Pearson	0.773	0.7536	0.747	0.7418	0.7374
CPC	0.7571	0.7405	0.7334	0.7289	0.7249
Jaccard	0.762	0.7432	0.7383	0.7353	0.7313
SPCC	0.7676	0.7509	0.7437	0.7386	0.7342
MSD	0.7298	0.7144	0.7101	0.7067	0.7026

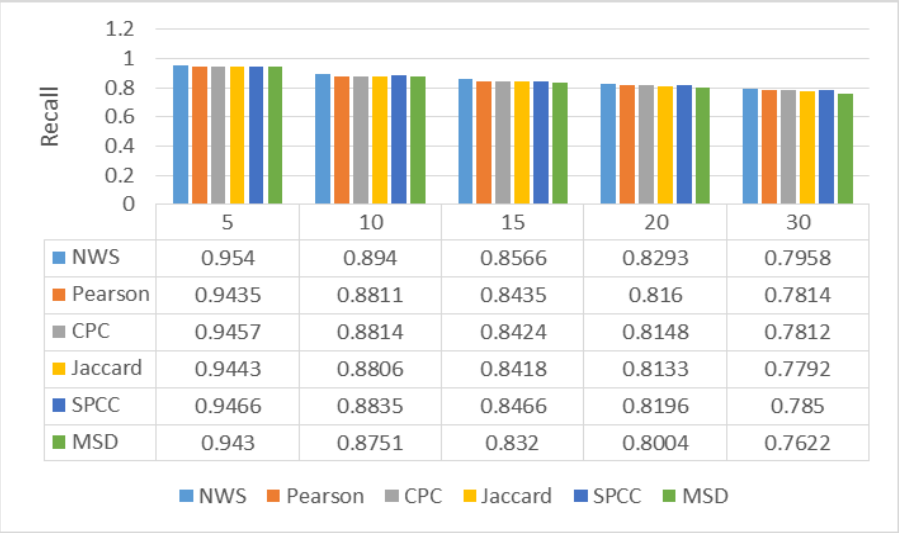
شکل ۶: صحت سیستم برای بازه‌سنی ۲۰ تا ۳۹ سال



	5	10	15	20	30
NWS	0.7703	0.7622	0.7592	0.7571	0.754
Pearson	0.7483	0.7375	0.7347	0.7326	0.7289
CPC	0.7556	0.7411	0.7397	0.7355	0.7324
Jaccard	0.7707	0.757	0.7524	0.7502	0.7471
SPCC	0.7561	0.7427	0.7381	0.7344	0.7315
MSD	0.7352	0.7269	0.722	0.7199	0.7186

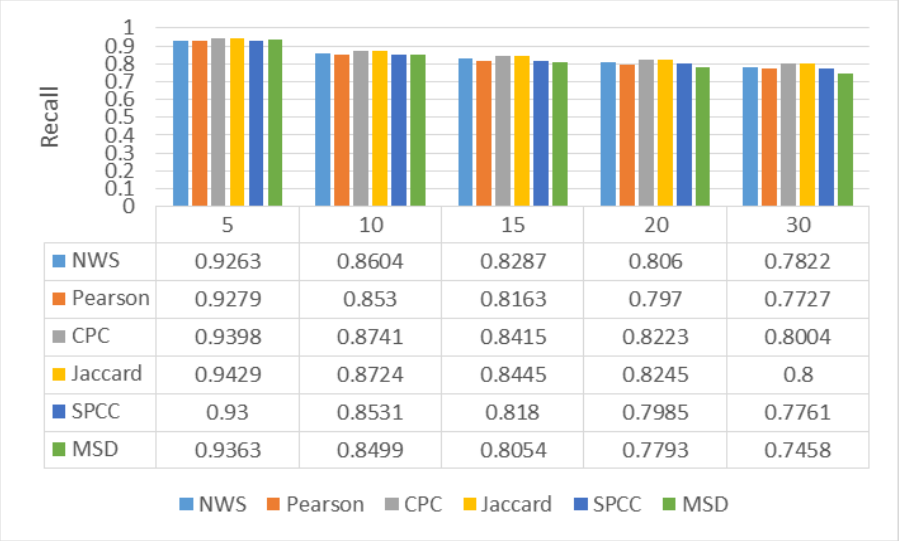
شکل ۷: صحت سیستم برای بازه‌سنی ۴۰ تا ۶۰ سال

ISSN: 2745-3673
http://www.Rcsj.ir



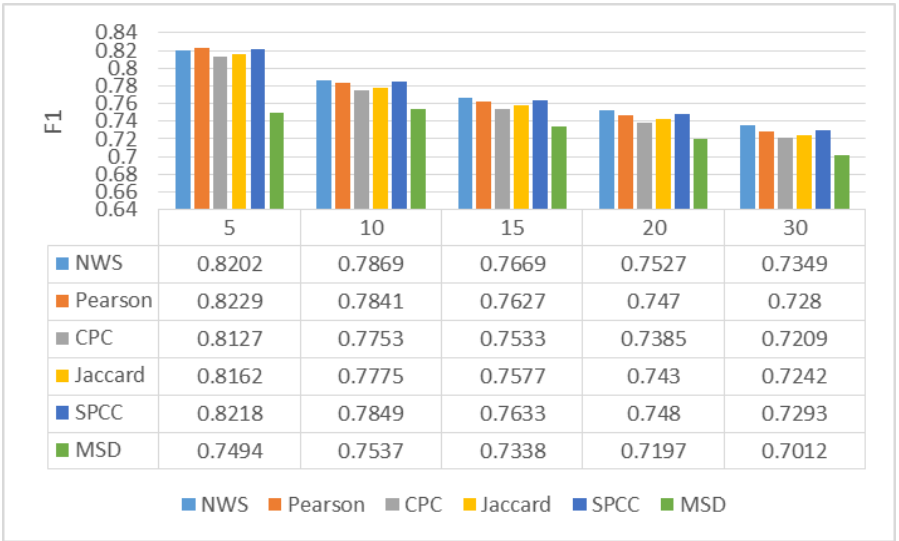
	5	10	15	20	30
NWS	0.954	0.894	0.8566	0.8293	0.7958
Pearson	0.9435	0.8811	0.8435	0.816	0.7814
CPC	0.9457	0.8814	0.8424	0.8148	0.7812
Jaccard	0.9443	0.8806	0.8418	0.8133	0.7792
SPCC	0.9466	0.8835	0.8466	0.8196	0.785
MSD	0.943	0.8751	0.832	0.8004	0.7622

شکل ۸: بازیابی سیستم برای بازه‌سنی ۲۰ تا ۳۹ سال



	5	10	15	20	30
NWS	0.9263	0.8604	0.8287	0.806	0.7822
Pearson	0.9279	0.853	0.8163	0.797	0.7727
CPC	0.9398	0.8741	0.8415	0.8223	0.8004
Jaccard	0.9429	0.8724	0.8445	0.8245	0.8
SPCC	0.93	0.8531	0.818	0.7985	0.7761
MSD	0.9363	0.8499	0.8054	0.7793	0.7458

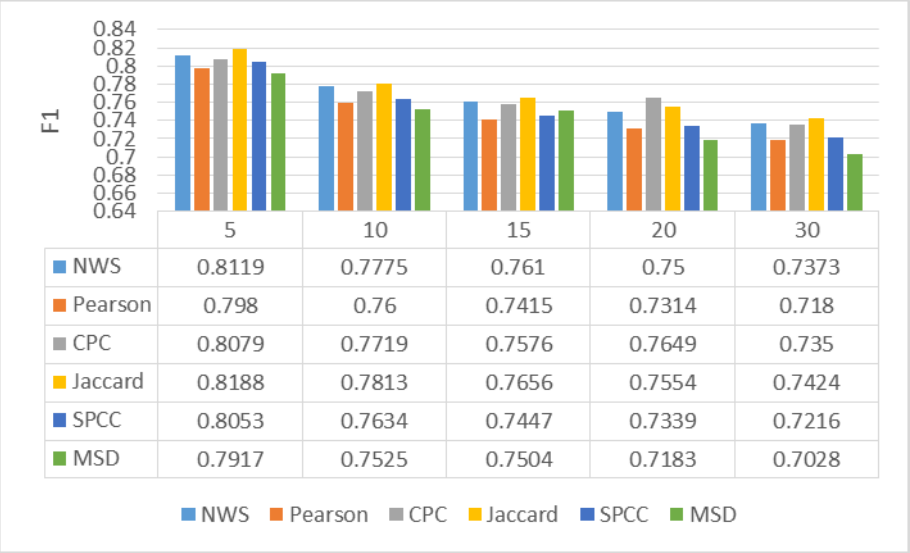
شکل ۹: بازیابی سیستم برای بازه‌سنی ۴۰ تا ۶۰ سال



	5	10	15	20	30
NWS	0.8202	0.7869	0.7669	0.7527	0.7349
Pearson	0.8229	0.7841	0.7627	0.747	0.728
CPC	0.8127	0.7753	0.7533	0.7385	0.7209
Jaccard	0.8162	0.7775	0.7577	0.743	0.7242
SPCC	0.8218	0.7849	0.7633	0.748	0.7293
MSD	0.7494	0.7537	0.7338	0.7197	0.7012

شکل ۱۰: معیار F1 سیستم برای بازه‌سنی ۲۰ تا ۳۹ سال

ISSN: 2745-3673
http://www.Rcsj.ir



	5	10	15	20	30
NWS	0.8119	0.7775	0.761	0.75	0.7373
Pearson	0.798	0.76	0.7415	0.7314	0.718
CPC	0.8079	0.7719	0.7576	0.7649	0.735
Jaccard	0.8188	0.7813	0.7656	0.7554	0.7424
SPCC	0.8053	0.7634	0.7447	0.7339	0.7216
MSD	0.7917	0.7525	0.7504	0.7183	0.7028

شکل ۱۱: معیار F1 سیستم برای بازه‌سنی ۴۰ تا ۶۰ سال

ISSN: 2745-3673
http://www.Rcsj.ir

نتیجه‌گیری

سیستم‌های توصیه‌گر به عنوان کاربردی هوشمند و تحت وب جهت تجارت الکترونیک که امروزه توسط وب‌سایت‌ها ارائه می‌گردد، بسیار کارا و سودمند هستند و با پیشنهاداتش به کاربران کمک می‌کند تا تصمیم بهتری در زمان کوتاه‌تر بگیرند و سود و فروش سایت را افزایش دهد. پالایش همکارانه یکی از مهم‌ترین رویکردهای استفاده شده در سیستم‌های توصیه‌گر است. این رویکرد از مشکلات مقیاس‌پذیری، تنکی‌داده و خطای پیش‌بینی بالا رنج می‌برد. هدف این مقاله، طراحی یک سیستم توصیه‌گر برای شخصی‌سازی پیشنهادات براساس اطلاعات سنی کاربران به همراه یک معیارشباهت وزن‌دار جدید با نام NWS است. دسته‌بندی کاربران براساس بازه‌های سنی، باعث بهبودپذیری مساله مقیاس‌پذیری گشته و تضعیف اثر مشکل تنکی‌داده با بدست آوردن درجه اطمینان تشابه کاربران توسط معیارشباهت جدید NWS، باعث بهبود خطای پیش‌بینی و عملکرد سیستم می‌شود. این سیستم با تفکیک کردن کاربران به دو بازه سنی ۲۰ تا ۳۹ سال و ۴۰ تا ۶۰ سال، و اهمیت دادن به کاربرانی که تعداد فیلم‌های مشترک زیادی با کاربرفعال دارند، توانست پیشنهادات را برای کاربران فعال شخصی‌سازی کند. نتایج آزمایشات نسبت به سایر روش‌های توصیه‌گری معمول، بیانگر بهبود نرخ خطا، عملکرد و کارایی سیستم است. حداکثر نرخ خطای بهبود یافته برای کاربران بین ۲۰ تا ۳۹ سال ۱۷,۴ درصد و برای کاربران بین ۴۰ تا ۶۰ سال ۲۰,۲ درصد، می‌باشد.

سپاسگزاری

بدینوسیله نویسنده از داوران ارجمند برای نظرات ارزنده و پیشنهاداتی در راستای بهبود ارائه این مقاله، تشکر می‌کند.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[1] Aggarwal C. Recommender systems. Springer International Publishing, 2016.

[۲] درزی، م.، مرادی‌منش، ز.، اصغری، ح.، "مقدمه‌ای بر سیستم‌های توصیه‌گر: الگوریتم‌ها و کاربردها".سازمان انتشارات جهاد دانشگاهی، تهران، ۱۳۸۹.

[۳] [۳]خلجی، م.، دادخواه، چ.، FNHSM_HRS": سیستم توصیه‌گر ترکیبی مبتنی بر خوشه‌بندی فازی و معیار شباهت اکتشافی،" هفتمین کنگره مشترک سیستم‌های فازی و هوشمند ایران، بجنورد، ص ۵۶۲-۵۶۸، بهمن ۱۳۹۷.

[4]Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J. (1994, October). GroupLens: an open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work(pp. 175-186). ACM.

[5]Koutrika, G., Bercovitz, B. and Garcia-Molina, H. (2009, June). FlexRecs: expressing and combining flexible recommendations. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (pp. 745-758). ACM.

[6]Jamali, M. and Ester, M. (2009, June). Trustwalker: a random walk model for combining trust-based and item-based recommendation. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 397-406). ACM.

[7]Shardanand, U. and Maes, P. (1995, May). Social information filtering: algorithms for automating" word of mouth". In Chi(Vol. 95, pp. 210-217.(

[8]Cacheda, F., Carneiro, V., Fernández, D. and Formoso, V. (2011). Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. ACM Transactions on the Web (TWEB), 5(1), 2.

[9]Goldberg D., Nichols D., Oki B. and Terry D. (1992). Using collaborative filtering to weave an information tapestry, Communications of the ACM, 35.12, 61-70.

[۱۰]خداوردی، ن.، دادخواه، چ.، خلجی، م.، "سامانه توصیه‌گر ترکیبی مبتنی بر خوشه بندی و محبوبیت،" کنفرانس بین المللی فناوری و نوآوری در علوم، مهندسی و تکنولوژی، ایران، تهران، دانشگاه شهید بهشتی، اسفند ۱۳۹۷.

[11] Ahn, H. J. (2008). A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. Information Sciences, 178(1), 37-51.

[12] Liu, H., Hu, Z., Mian, A., Tian, H. and Zhu, X. (2014). A new user similarity model to improve the accuracy of collaborative filtering. Knowledge-Based Systems, 56, 156-166.

[13] Herlocker, J., Konstan, J. A. and Riedl, J. (2002). An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. Information retrieval, 5(4), 287-310.

[14] McLaughlin, M. R. and Herlocker, J. L. (2004, July). A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 329-336). ACM.

[15] Bellogín, A., Castells, P. and Cantador, I. (2013, May). Improving memory-based collaborative filtering by neighbour selection based on user preference overlap. In Proceedings of the 10th conference on open research areas in information retrieval (pp. 145-148). LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

[16] Choi, K. and Suh, Y. (2013). A new similarity function for selecting neighbors for each target item in collaborative filtering. Knowledge-Based Systems, 37, 146-153.

ISSN: 2745-3673
http://www.Rcsj.ir

[17] Javari, A., Gharibshah, J. and Jalili, M. (2014). Recommender systems based on collaborative filtering and resource allocation. Social Network Analysis and Mining, 4(1), 234.

[18] Zhang, J., Lin, Y., Lin, M. and Liu, J. (2016). An effective collaborative filtering algorithm based on user preference clustering. Applied Intelligence, 45(2), 230-240.

[19] Park, Y., Park, S., Jung, W. and Lee, S. G. (2015). Reversed CF: A fast collaborative filtering algorithm using a k-nearest neighbor graph. Expert Systems with Applications, 42(8), 4022-4028.

[۲۰]خلجی، م.، بهبود عملکرد سیستم پیشنهاددهنده ترکیبی با استفاده از شبکه عصبی و تخصیص منابع، پایان‌نامه کارشناسی ارشد هوش مصنوعی و رباتیکز، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی خواجه نصیرالدین طوسی، ایران، تهران، شهریور ۱۳۹۶.

[21] Koohi, H., and Kiani, K. (2016). User based Collaborative Filtering using fuzzy C-means. Measurement, 91, 134-139.

[22] Belacel, N., Durand, G., Leger, S. and Bouchard, C. (2018, January). Scalable Collaborative Filtering Based on Splitting-Merging Clustering Algorithm. In International Conference on Agents and Artificial Intelligence (pp. 290-311). Springer, Cham.

[23] Kant, S., Mahara, T., Jain, V. K., Jain, D. K. and Sangaiah, A. K. (2018). LeaderRank based k-means clustering initialization method for collaborative filtering. Computers & Electrical Engineering, 69, 598-609.

[24] Khalaji, M. and Mohammadnejad N. (2019). FCNHSMRA_HRS: Improve the performance of the movie hybrid recommender system using resource allocation approach, In Proceedings of the 4th International Conference on Researchers in Science & Engineering & International Congress on Civil, Architecture and Urbanism in Asia, Kasem Bundit University, Bangkok, Thailand.

[25] Wang, Y., Feng, C., Guo, C., Chu, Y. and Hwang, J. N. (2019, January). Solving the Sparsity Problem in Recommendations via Cross-Domain Item Embedding Based on Co-Clustering. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (pp. 717-725). ACM.