ارائه یک روش ترکیبی جدید بر اساس تکنیک گسترش پروفایل برای حل مسئله شروع سرد در سیستم‌های توصیه‌گر

فریاد طهماسبی[۱]، دانشجوی کارشناسی ارشد؛ مجید مقدادی[۲]، استادیار؛ سجاد احمدیان[۳]، دانشجوی دکتری

۱- دانشکده مهندسی - دانشگاه زنجان - زنجان - ایران – faryadtahmasebi@znu.ac.ir
۲- دانشکده مهندسی - دانشگاه زنجان - زنجان - ایران – meghdadi@znu.ac.ir
۳- دانشکده مهندسی - دانشگاه زنجان - زنجان - ایران – s.ahmadian@znu.ac.ir

چکیده: با توجه به رشد روزافزون حجم اطلاعات در دسترس در محیط جهانی وب، فرآیند جستجوی اطلاعات مورد نیاز کاربران با صرف زمان زیادی صورت می‌گیرد. سیستم‌های توصیه‌گر به‌منظور جلوگیری از اتلاف وقت کاربران، اطلاعاتی را در اختیار آن‌ها قرار می‌دهند که به احتمال زیاد مفید و ارزشمند هستند. متداول‌ترین روشی که برای تولید پیشنهاد در سیستم‌های توصیه‌گر مورد استفاده قرار می‌گیرد، روش پالایش مشارکتی است. این روش با وجود استفاده زیاد، دارای مشکلاتی نظیر مشکل شروع سرد است. در این مقاله یک روش ترکیبی جدید بر اساس تکنیک گسترش پروفایل برای بهبود مشکل شروع سرد ارائه شده است. در روش پیشنهادی علاوه بر شباهت رتبه‌دهی کاربران از شباهت دموگرافی کاربران به‌منظور انتخاب مجموعه همسایگان قوی‌تر برای کاربر هدف استفاده شده است. سیستم پیشنهادی با استفاده از مجموعه داده‌ای MovieLens مورد ارزیابی قرار گرفته و نتایج حاصل از آن حاکی از بهبود کارایی سیستم توصیه‌گر پیشنهادی نسبت به سایر روش‌ها است.

واژه‌های کلیدی: سیستم‌های توصیه‌گر، پالایش مشارکتی، شروع سرد، گسترش پروفایل، معیار شباهت، اطلاعات دموگرافی.

A Novel Hybrid Approach based on Profile Expansion Technique to Improve Cold Start Problem in Recommender Systems

F. Tahmasebi, MSc Student[1]; M. Meghdadi, Assistant Professor[2]; S. Ahmadian, PhD Student[3]

1- Faculty of Engineering, University of Zanjan, Zanjan, Iran, Email: faryadtahmasebi@znu.ac.ir
2- Faculty of Engineering, University of Zanjan, Zanjan, Iran, Email: meghdadi@znu.ac.ir
3- Faculty of Engineering, University of Zanjan, Zanjan, Iran, Email: s.ahmadian@znu.ac.ir

Abstract: Due to the growing volume of information available on the Web, the data search process is performed with spending a lot of time. In order to avoid wasting time of users, recommender systems provide information for them which is likely to be useful and valuable. Collaborative filtering is the most popular approach to provide recommendations for the users in recommender systems. However, it suffers from some problems such as cold start problem. In this paper, we present a novel hybrid approach based on profile expansion technique to improve the cold start problem in the recommender systems. In the proposed method, we take into consideration user's demographic data beside user's rating data in order to find an enrich neighborhood set for the active user. The results of experiments on MovieLens dataset showed that the proposed method outperformed the other recommendation methods.

Keywords: Recommender systems, collaborative filtering, cold start, profile expansion, similarity metric, demographic data.

تاریخ ارسال مقاله: ۱۳۹۵/۰۶/۱۰
تاریخ اصلاح مقاله: ۱۳۹۵/۰۹/۰۵، ۱۳۹۵/۱۰/۱۴، ۱۳۹۵/۱۱/۱۲ و ۱۳۹۵/۱۱/۲۸
تاریخ پذیرش مقاله: ۱۳۹۵/۱۲/۰۹
نام نویسنده مسئول: مجید مقدادی
نشانی نویسنده مسئول: ایران- زنجان-دانشگاه زنجان-دانشکده مهندسی- گروه مهندسی کامپیوتر.

۱- مقدمه

سیستم‌های توصیه‌گر[۱]، به‌عنوان زیرمجموعه‌ای از سیستم‌های فیلترسازی اطلاعات[۲] شناخته می‌شوند که به پیش‌بینی رتبه کاربران برای مجموعه‌ای از آیتم‌ها (کتاب‌ها، فیلم‌ها، اخبار، آهنگ‌ها و...) می‌پردازند [۱]. سیستم‌های توصیه‌گر به کاربران کمک می‌کنند که آیتم‌های مورد علاقه خود را از میان هزاران آیتم موجود پیدا کنند. درواقع، دلیل موفقیت سیستم‌های توصیه‌گر در وب‌سایت‌های تجاری، شخصی‌سازی کردن[۳] پیشنهادها به کاربران است [۲، ۳].

روش پالایش مشارکتی یکی از پراستفاده‌ترین و رایج‌ترین روش‌های تولید پیشنهاد در سیستم‌های توصیه‌گر است [۴]. در این روش کاربران مشابه (کاربران دارای رتبه‌دهی مشترک) با کاربر هدف، تحت عنوان مجموعه همسایگان کاربر هدف شناخته می‌شوند. کاربر هدف به کاربری گفته می‌شود که سیستم توصیه‌گر به تولید پیشنهاد برای او می‌پردازد. بنابراین، سیستم توصیه‌گر مبتنی بر پالایش مشارکتی لیستی از N آیتم مورد علاقه از میان آیتم‌های رتبه‌دهی‌شده توسط این مجموعه همسایگان را به کاربر هدف پیشنهاد می‌دهد. روش پالایش مشارکتی به دو دسته مبتنی بر حافظه و مبتنی بر مدل تقسیم می‌شود [۵]. سیستم‌های پالایش مشارکتی مبتنی بر حافظه [۶]، به‌صورت مستقیم از ماتریس کاربر- آیتم به‌منظور محاسبه پیش‌بینی و تولید پیشنهاد برای کاربر هدف استفاده می‌کنند. این روش‌ها از تکنیک‌های آماری نظیر شباهت پیرسون [۷] و شباهت کسینوسی [۸] به‌منظور محاسبه شباهت بین کاربران و پیداکردن مجموعه همسایگان کاربر هدف استفاده می‌کنند. در راهکار پالایش مشارکتی مبتنی بر مدل [۹]، با استفاده از تکنیک‌های یادگیری ماشین، نظیر خوشه‌بندی [۱۰]، طبقه‌بند بیزین [۱۱] و الگوریتم ژنتیک [۱۲] یک مدل از داده‌های آموزشی موجود در ماتریس کاربر- آیتم ساخته می‌شود. علاوه‌براین، در مرحله تست، از این مدل آموزشی به‌منظور پیش‌بینی رتبه کاربران استفاده می‌شود.

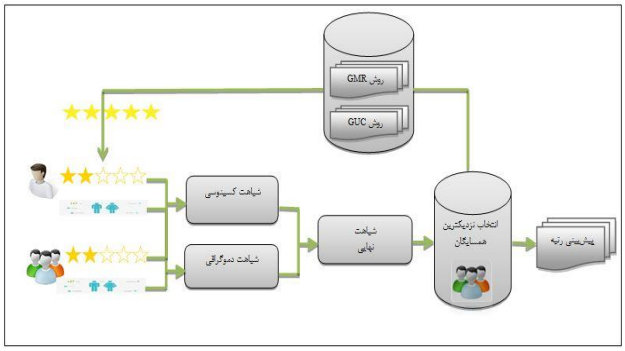
یکی از مهم‌ترین مسائل در حوزه سیستم‌های توصیه‌گر مشکل شروع سرد است. مشکل شروع سرد به دلیل تعداد کم آیتم‌های رتبه‌دهی‌شده توسط کاربران رخ می‌دهد. مسئله شروع سرد خود به دو دسته مسئله کاربر جدید و آیتم جدید تقسیم می‌شود [۱۳]. تمرکز اصلی ما در این مقاله بر روی مسئله کاربر جدید است. مسئله کاربر جدید هنگامی رخ می‌دهد که کاربر به‌تازگی وارد سیستم شده و هیچ آیتمی را رتبه‌دهی نکرده است و یا از قبل در سیستم حضور داشته، ولی کم‌تر فعال بوده و آیتم‌های کمی را رتبه‌دهی کرده است [۱۴].

تاکنون کارهای زیادی به‌منظور حل مشکل شروع سرد در سیستم‌های توصیه‌گر ارائه شده است. این روش‌ها در حالت کلی به دو دسته تقسیم می‌شوند [۲، ۱۴]. گروهی از روش‌ها، روش‌های ترکیبی می‌باشند. کاربران جدید تعداد آیتم کمی را رتبه‌دهی کرده‌اند و دارای یک پروفایل خالی یا یک پروفایل خیلی کوچک می‌باشند. بنابراین سیستم‌های توصیه‌گر قادر به تشخیص ترجیحات این کاربران نیستند و نمی‌توانند به این کاربران آیتم‌های مناسب و مورد علاقه آن‌ها را پیشنهاد دهند. ازاین‌رو، بیشتر محققان از روش‌های ترکیبی به‌منظور حل مسئله کاربر جدید استفاده کرده‌اند. روش‌های ترکیبی، معمولاً ترکیبی از روش پالایش مشارکتی با سایر منابع داده اضافی می‌باشند [۱۵].

تاکنون تحقیقات زیادی در مورد استفاده از منابع داده اضافی نظیر اطلاعات دموگرافی کاربران به‌منظور حل مشکل شروع سرد انجام شده است. در مرجع [۱۶]، یک روش جدید برای حل مسئله شروع سرد که شامل سه فاز مختلف است، مورد بررسی قرار گرفته است. در فاز اول با استفاده از الگوریتم‌های طبقه‌بندی، کاربر جدید در یک گروه خاص قرار می‌گیرد. در فاز دوم، شباهت بین کاربر جدید و دیگر کاربران موجود در گروه مدنظر، بر اساس اطلاعات دموگرافی کاربران محاسبه می‌شود. در فاز سوم، با استفاده از گروه‌های مختلف ایجادشده برای کاربران، پیش‌بینی رتبه برای کاربر جدید انجام می‌گیرد. در مرجع [۱۷]، یک روش ترکیبی برای حل مشکل شروع سرد با استفاده از اطلاعات مبتنی بر محتوا و اطلاعات اجتماعی کاربران ارائه شده است. ایده اصلی این روش ساخت اطلاعات محتوایی بر اساس مشخصات کلمات کلیدی مرتبط با آیتم‌های مختلف و استفاده از این اطلاعات به‌منظور تولید پیشنهاد برای کاربران شروع سرد است. در [۱۸] روش‌های مختلفی به‌منظور پروفایل‌سازی برای کاربران بر اساس اطلاعات دموگرافیک بررسی شده‌اند. این روش‌ها ترکیبات مختلفی را در نظر می‌گیرند که شامل انواع ویژگی‌های مورد استفاده، نحوه ارائه ویژگی‌ها و راه‌های پروفایل‌سازی کاربران می‌باشند. در مرجع [۱۹] یک روش به‌منظور حل مشکل شروع سرد کاربران بر اساس اطلاعات دموگرافی ارائه شده است. در این روش، به‌جای استفاده از رتبه‌های داده‌شده به آیتم‌های مختلف توسط کاربران شروع سرد، از اطلاعات دموگرافی آن‌ها به‌منظور تولید پیشنهادهایی به این‌گونه کاربران استفاده شده است. بدین منظور، یک چارچوب برای ارزیابی ویژگی‌های مختلف دموگرافی نظیر سن، جنسیت و شغل پیشنهاد شده است. در [۲۰] یک سیستم توصیه‌گر مبتنی بر اطلاعات دموگرافی کاربران به‌منظور حل مشکل شروع سرد ارائه شده است. در این روش، کاربران مختلف بر اساس اطلاعات دموگرافی آن‌ها دسته‌بندی می‌شوند و سپس بر اساس دسته‌بندی دموگرافی آن‌ها، توصیه‌هایی تولید می‌شوند. بنابراین، در این روش با استفاده از اطلاعات دموگرافی کاربران، مشکل شروع سرد تا حد زیادی حل می‌شود.

گروهی دیگر از روش‌ها، برخلاف روش‌های ترکیبی برای حل مسئله کاربر جدید از منابع داده اضافی استفاده نکرده‌اند و تنها با در نظر گرفتن وضعیت فعلی پروفایل رتبه‌دهی کاربران مسئله شروع سرد را بهبود داده‌اند. در مرجع [۲۱]، از یک چارچوبی که بر مبنای معیار شباهت محلی کاربران و معیار شباهت سراسری کاربران است، برای بهبود مشکل شروع سرد استفاده شده است. علاوه‌براین، انواع پراکندگی داده نظیر پراکندگی کلی داده[۴]، پراکندگی خاص کاربر[۵]، پراکندگی خاص کاربر- آیتم[۶]، و معیار پراکندگی یکنواخت[۷]، برای حل مسئله پراکندگی داده و مسئله شروع سرد استفاده شده است. در این

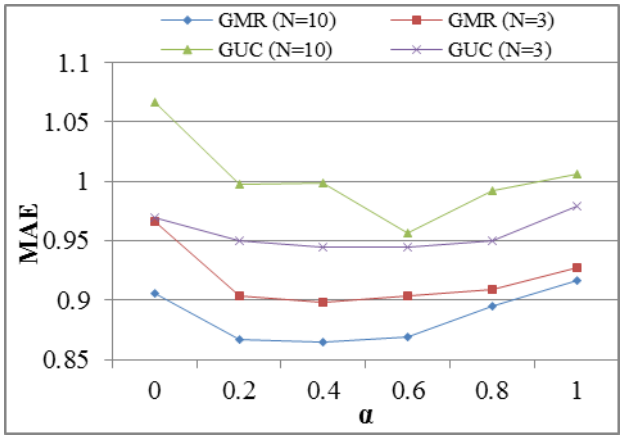
روش، به‌منظور پیش‌بینی رتبه یک آیتم خاص برای کاربر هدف از ترکیب رتبه‌های تولیدشده بر اساس شباهت‌های محلی و سراسری کاربران و با در نظر گرفتن مقدار پراکندگی یکنواخت ماتریس رتبه‌دهی کاربران استفاده شده است. مرجع [۲۲]، به محدودیت‌های معیارهای شباهت سنتی نظیر معیار شباهت کسینوسی و پیرسون پرداخته است، و یک معیار شباهت اکتشافی جدید بنام PIP[۸]، برای مقابله با مشکل شروع سرد ارائه داده است. این معیار شباهت شامل سه فاکتور Proximity (فاصله بین دو رتبه را محاسبه می‌کند)، فاکتور Impact (شدت تنفر یا علاقه کاربران به آیتم مورد نظر را نشان می‌دهد) و فاکتور Popularity (فاصله میانگین رتبه‌دهی دو کاربر به آیتم مدنظر از میانگین رتبه‌دهی کل کاربران به آیتم مدنظر را محاسبه می‌کند) است.



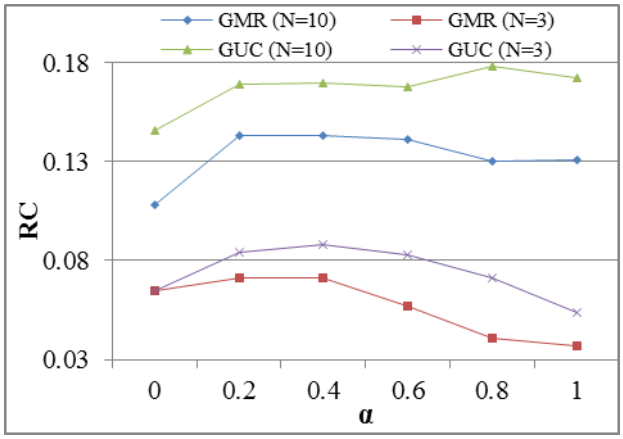
شکل ۱: چارچوب کلی روش پیشنهادی.

مرجع [۲]، به‌منظور مقابله با مشکل انتخاب مجموعه همسایگان ضعیف برای کاربر جدید در روش پالایش مشارکتی، اقدام به گسترش پروفایل کاربر جدید بر اساس تکنیک‌های آیتم سراسری[۹]، آیتم محلی[۱۰] و کاربر محلی[۱۱] نموده است. تکنیک آیتم سراسری تلاش می‌کند مجموعه‌ای از شبیه‌ترین آیتم‌ها به آیتم‌هایی که از قبل در پروفایل کاربر موجود است را پیدا کند و به پروفایل کاربر اضافه نماید. تکنیک آیتم محلی شامل دو مرحله است، در مرحله اول، با توجه به پروفایل فعلی کاربر، سیستم توصیه‌گر اقدام به تولید پیشنهاد آیتم برای وی می‌کند و آیتم‌های با بالاترین درجه تخمین را به پروفایل کاربر اضافه می‌کند. در مرحله دوم، سیستم اقدام به تولید پیشنهاد آیتم‌ها به‌منظور ارائه دادن آن‌ها به کاربر جدید می‌کند. تکنیک کاربر محلی بر اساس مجموعه همسایگان جاری کاربر هدف اقدام به گسترش پروفایل می‌کند. در این روش از میان آیتم‌های رتبه‌دهی‌شده توسط مجموعه همسایگان کاربر هدف، تعدادی از آن‌ها بر اساس استراتژی‌هایی نظیر بیش‌ترین رتبه محلی[۱۲] (LMR) و خوشه‌بندی کاربر محلی[۱۳] (LUC) انتخاب می‌شوند.

در این مقاله، یک روش ترکیبی جدید مبتنی بر تکنیک گسترش پروفایل برای حل مسئله کاربر جدید ارائه شده است. در روش پیشنهادی، برخلاف روش ارائه‌شده در [۲] که تنها از اطلاعات مربوط به ماتریس رتبه‌های داده‌شده به آیتم‌ها توسط کاربران استفاده می‌کند، از اطلاعات دموگرافی کاربران نیز در کنار ماتریس رتبه‌ها استفاده شده است. درواقع، در روش پیشنهادی ترکیبی از شباهت کسینوسی بر اساس ماتریس رتبه‌ها و شباهت دموگرافی بر اساس اطلاعات دموگرافی کاربران به‌عنوان شباهت نهایی در مرحله گسترش پروفایل کاربران استفاده شده است.



(الف)

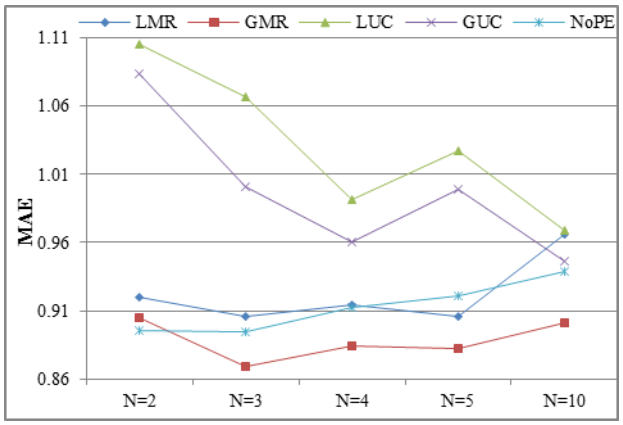


(ب)

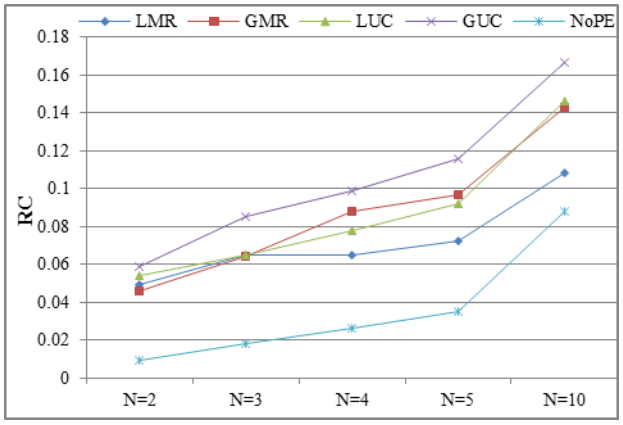
شکل ۲: تأثیر مقادیر مختلف پارامتر α بر روی روش پیشنهادی برای (الف) معیار MAE، (ب) معیار RC.

۲- روش پیشنهادی

در این بخش یک روش جدید به‌منظور حل مشکل شروع سرد در سیستم‌های توصیه‌گر ارائه شده است. در سیستم پیشنهادی فرآیند تولید پیشنهاد شامل دو فاز است. در فاز اول پروفایل کاربر هدف گسترش می‌یابد. بدین منظور از ترکیب شباهت کسینوسی و شباهت دموگرافی به‌عنوان شباهت نهایی برای انتخاب نزدیک‌ترین همسایگان کاربر هدف و گسترش پروفایل وی استفاده می‌شود. بنابراین، در روش پیشنهادی علاوه بر اطلاعات مربوط به رتبه‌های داده‌شده به آیتم‌های مختلف توسط کاربران، از اطلاعات دموگرافی آن‌ها نیز به‌عنوان اطلاعات اضافی برای حل مشکل شروع سرد استفاده شده است. پس از گسترش پروفایل کاربران، در فاز دوم رتبه آیتم‌های مربوط به کاربر هدف بر اساس پروفایل گسترش‌یافته پیش‌بینی می‌شود. در ادامه به‌شرح هر یک از فازها پرداخته می‌شود. شکل ۱ چارچوب کلی روش پیشنهادی را نشان می‌دهد.



شکل ۳: مقدار MAE برای I=۱۰ و مقادیر مختلف N.



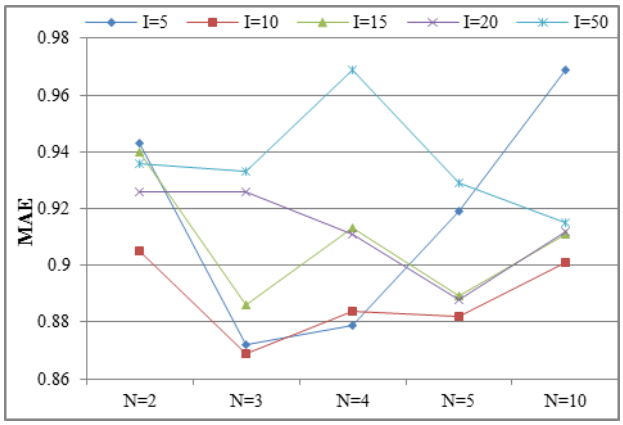
شکل ۴: مقدار RC برای I=۱۰ و مقادیر مختلف N.

۲-۱- فاز گسترش پروفایل

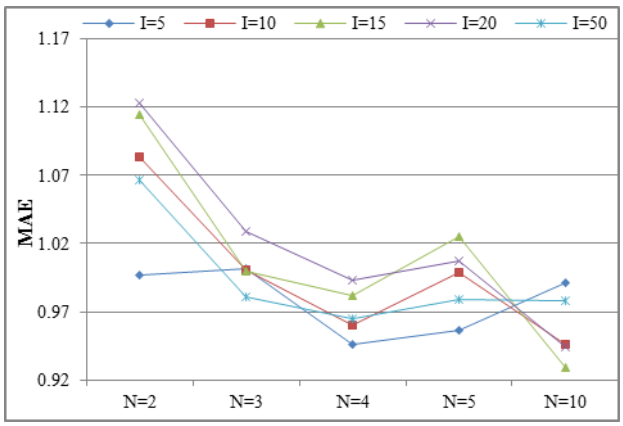
در این مرحله، یک روش ترکیبی به‌منظور گسترش پروفایل کاربران ارائه شده است که بر اساس شباهت‌های کسینوسی و دموگرافی بین کاربران است. از اطلاعات موجود در ماتریس رتبه‌های کاربران به آیتم‌های مختلف برای محاسبه شباهت کسینوسی و همچنین از اطلاعات دموگرافی کاربران (سن، جنسیت و شغل) برای محاسبه شباهت دموگرافی بین کاربران استفاده شده است. به‌منظور تحلیل مشکل شروع سرد، از روش پیشنهادشده در مقاله [۲] استفاده شده است. در این روش، برای هرکدام از کاربران مورد ارزیابی تعداد N آیتم تصادفی انتخاب می‌شود که مقدار N کم‌تر از تعداد کل رتبه‌های داده‌شده توسط کاربر است. سپس برای محاسبه شباهت رتبه‌دهی کاربران از معیار شباهت کسینوسی مطابق رابطه (۱) استفاده می‌شود [۲].

$$sim(u,v)_{\cos} = \frac{\sum_{i\in A_{u,v}} r_{u,i} r_{v,i}}{\sqrt{\sum_{i\in A_{u,v}} r_{u,i}^{\,2} \sum_{i\in A_{u,v}} r_{v,i}^{\,2}}} \qquad (۱)$$

در رابطه فوق $sim(u,v)_{cos}$ شباهت کسینوسی بین دو کاربر u و v است که بازه آن بین ۰ و ۱ است، $r_{u,i}$ رتبه‌ای است که کاربر u به آیتم i داده است، و $A_{u,v}$ مجموعه آیتم‌های مشترک بین دو کاربر u و v است.



(الف)



(ب)

شکل ۵: مقدار MAE به‌ازای مقادیر مختلف پارامترهای N و I برای (الف) روش GMR، (ب) روش GUC.

در ادامه شباهت کاربران بر اساس داده‌های دموگرافی (سن، جنسیت و شغل) نیز محاسبه می‌شود. شباهت دموگرافی بین دو کاربر مطابق رابطه (۲) محاسبه می‌شود. در این رابطه اگر مقدار یک صفت خاص از دو کاربر یکسان باشند، مقدار شباهت این صفت خاص برای دو کاربر برابر یک و در غیراین‌صورت برابر صفر است. به‌عنوان مثال، اگر هر دو کاربر دارای جنسیت یکسان باشند، مقدار شباهت یک، در غیراین‌صورت مقدار شباهت صفر است [۱۶].

$$sim(u,v)_{demo} = \frac{\sum_{j=1}^{l} s_j w_j}{\sum_{j=1}^{l} w_j} \qquad (۲)$$

در این رابطه $sim(u,v)_{demo}$ شباهت دموگرافی بین دو کاربر u و v است که بازه آن بین ۰ و ۱ است. همچنین، s_j مقدار شباهت بین jامین صفت از دو کاربر و w_j وزن متناظر jامین صفت است.

بعد از محاسبه شباهت‌های کسینوسی و دموگرافی بین کاربران، برای محاسبه شباهت نهایی از رابطه (۳) به‌صورت زیر استفاده می‌شود:

$$sim(u,v) = (1-\alpha) sim(u,v)_{\cos} + \alpha sim(u,v)_{demo} \qquad (۳)$$

در رابطه (۳) پارامتر α تعیین‌کننده وابستگی شباهت نهایی بین کاربران به هر یک از معیارهای شباهت کسینوسی و دموگرافی است. در این رابطه مقدار α=۱ نشان‌دهنده وابستگی کامل شباهت نهایی به

شباهت دموگرافی کاربران است، و مقدار α=۰ نشان‌دهنده وابستگی کامل شباهت نهایی به شباهت کسینوسی کاربران است. با توجه به رابطه (۳) می‌توان نتیجه گرفت که حتی اگر یک کاربر هیچ آیتم رتبه‌داده‌شده‌ای نداشته باشد، شباهت آن با سایر کاربران بر اساس معیار شباهت دموگرافی قابل محاسبه است.

پس از محاسبه شباهت نهایی بین کاربران، تعدادی از کاربران که دارای بیش‌ترین میزان شباهت با کاربر هدف می‌باشند، به‌عنوان مجموعه نزدیک‌ترین همسایگان کاربر هدف انتخاب می‌شوند. سپس تعداد مشخصی از آیتم‌های رتبه‌دهی شده توسط این مجموعه از نزدیک‌ترین همسایگان کاربر هدف بر اساس یکی از استراتژی‌های GMR[۱۴] یا GUC[۱۵] (که در ادامه شرح داده می‌شوند) و طبق رابطه (۴) به پروفایل کاربر اضافه می‌شوند.

$$pre(u,i)=\frac{\sum_{v\in N_u} sim(u,v)\quad r_{v,i}}{\sum_{v\in N_u} sim(u,v)} \qquad (۴)$$

$pre(u,i)$ رتبه پیش‌بینی‌شده برای کاربر هدف u در مورد آیتم i است، $sim(u,v)$ شباهت بین دو کاربر u و v است که با استفاده از رابطه (۳) محاسبه می‌شود، N_u مجموعه همسایگان کاربر هدف u است و $r_{v,i}$ رتبه داده شده به آیتم i توسط کاربر v است.

۲-۱-۱- استراتژی GMR

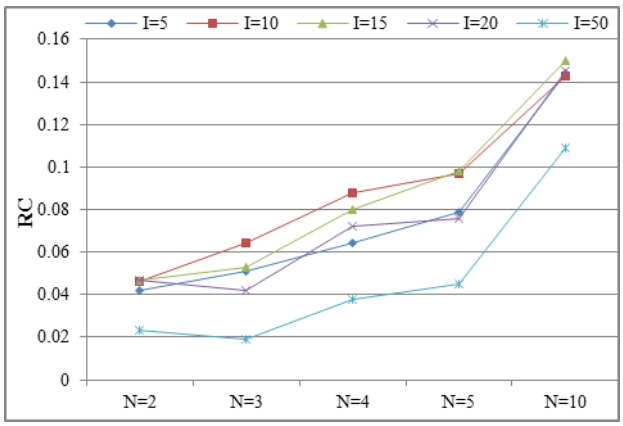
در این استراتژی، تعداد I آیتم مختلف به‌منظور گسترش پروفایل کاربر هدف انتخاب می‌شوند. بدین‌منظور، آیتم‌هایی انتخاب می‌شوند که دارای بیش‌ترین تعداد رتبه توسط کاربران موجود در همسایگی کاربر هدف باشند. بنابراین، در این استراتژی، آیتم‌های موجود بر اساس تعداد رتبه‌هایی که دریافت کرده‌اند به‌صورت نزولی مرتب شده و تعداد I آیتم از ابتدای این لیست برای گسترش پروفایل کاربر هدف انتخاب می‌شوند.

۲-۱-۲- استراتژی GUC

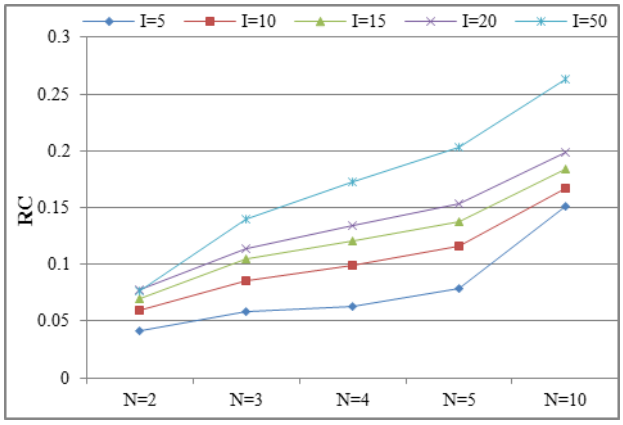
در این استراتژی شباهت بین آیتم‌های موجود در پروفایل کاربر هدف با دیگر آیتم‌های موجود در سیستم محاسبه می‌شود. سپس تعدادی از شبیه‌ترین آیتم‌ها به آیتم‌های موجود در پروفایل کاربر هدف به‌منظور فرآیند گسترش پروفایل انتخاب می‌شوند.

۲-۲- فاز پیش‌بینی رتبه

در این فاز بعد از این‌که پروفایل اولیه کاربران گسترش یافت، شباهت بین کاربر هدف و دیگر کاربران موجود در سیستم بر اساس پروفایل گسترش‌یافته آن‌ها مطابق با رابطه (۱) محاسبه می‌شود، و تعداد k از شبیه‌ترین کاربران به‌عنوان مجموعه همسایگان کاربر هدف انتخاب می‌شوند. درنهایت، با استفاده از این همسایگی به‌دست‌آمده برای کاربر هدف، رتبه کاربر به یک آیتم مشخص بر اساس رابطه (۴) تخمین زده می‌شود.

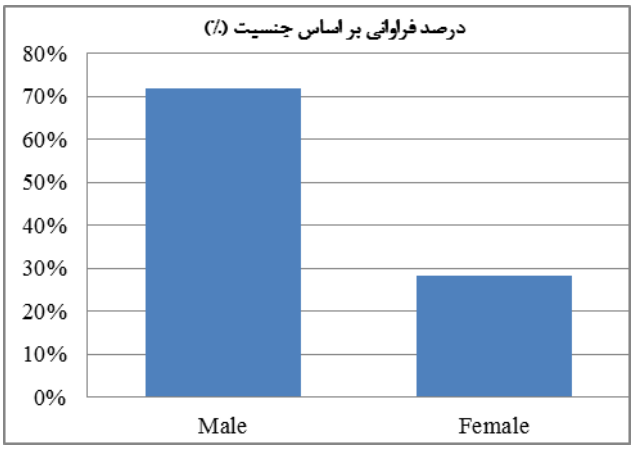


(الف)

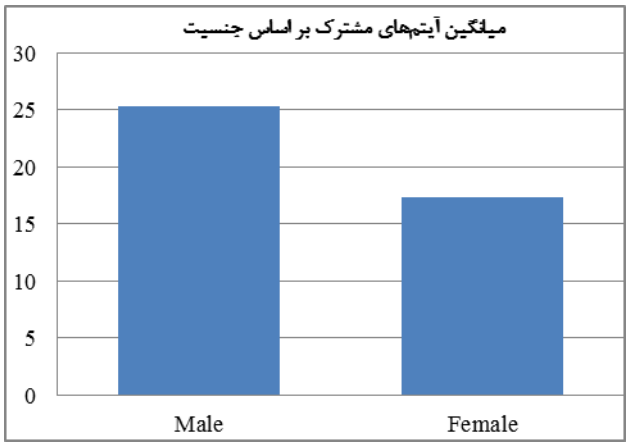


(ب)

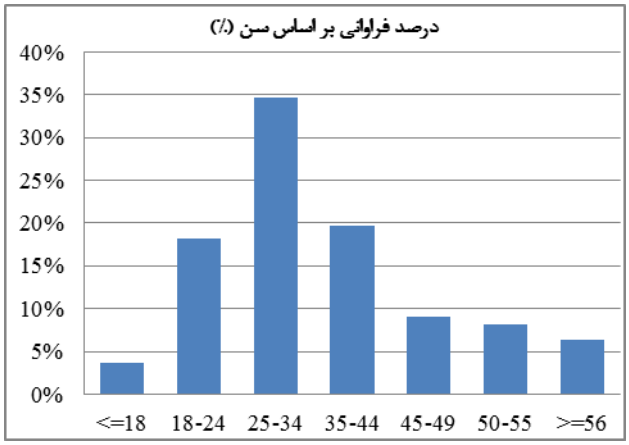
شکل ۶: مقدار RC به ازای مقادیر مختلف پارامترهای N و I برای (الف) روش GMR، (ب) روش GUC.



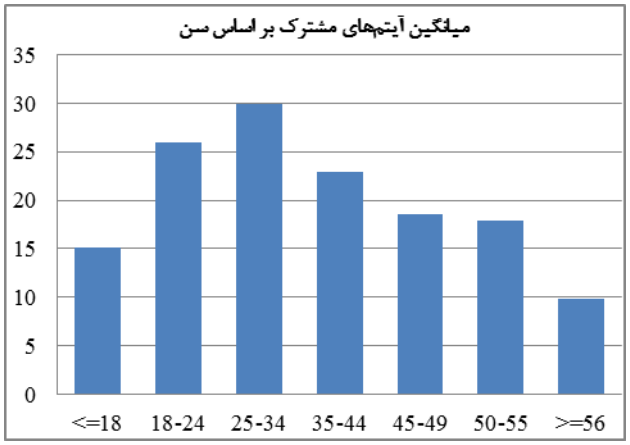
(الف)



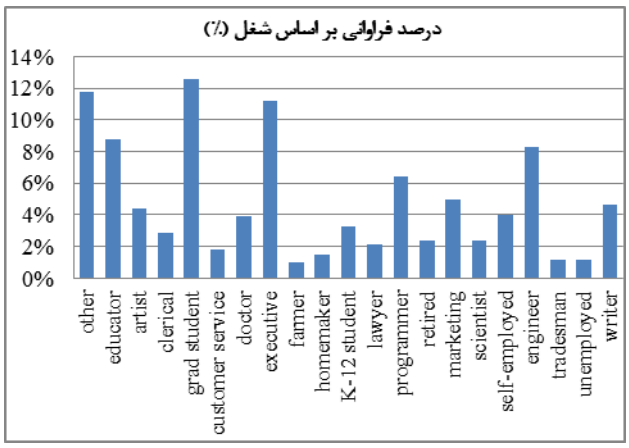
(ب)



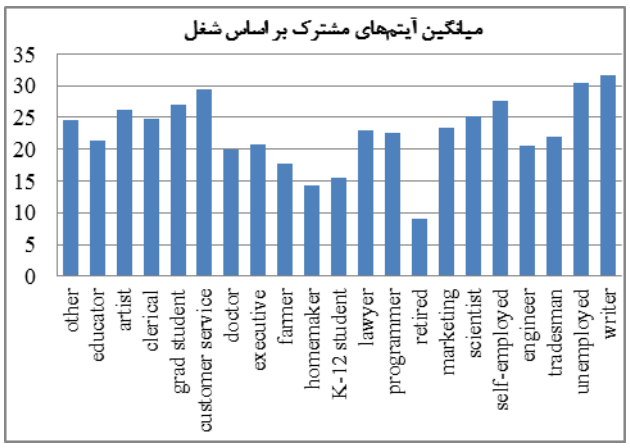
(ج)



(د)



(ه)



(ی)

شکل ۷: بررسی آماری مجموعه داده‌ای بر اساس اطلاعات دموگرافی کاربران: (الف) درصد فراوانی بر اساس جنسیت افراد، (ب) میانگین آیتم‌های مشترک بر اساس جنسیت افراد، (ج) درصد فراوانی بر اساس سن افراد، (د) میانگین آیتم‌های مشترک بر اساس سن افراد، (ه) درصد فراوانی بر اساس شغل افراد، (ی) میانگین آیتم‌های مشترک بر اساس شغل افراد.

جدول ۱: مقدار MAE و RC برای N=۳ و مقادیر مختلف I.

	Metric	LMR	GMR	LUC	GUC
I=۵	MAE	۰/۹۱۴	۰/۸۷۲	۱/۰۵۴	۱/۰۰۲
	RC	۰/۰۵۶	۰/۰۵۱	۰/۰۴۵	۰/۰۵۸
I=۱۰	MAE	۰/۹۰۶	۰/۸۶۹	۱/۰۶۷	۱/۰۰۱
	RC	۰/۰۶۵	۰/۰۶۴	۰/۰۶۵	۰/۰۸۵
I=۱۵	MAE	۰/۹۱۵	۰/۸۸۶	۱/۰۹۱	۱/۰۰۰
	RC	۰/۰۵۱	۰/۰۵۳	۰/۰۷۱	۰/۱۰۵
I=۲۰	MAE	۱/۰۱۶	۰/۹۲۶	۱/۱۴۳	۱/۰۲۹
	RC	۰/۰۵۰	۰/۰۴۲	۰/۰۷۷	۰/۱۱۴
I=۵۰	MAE	۰/۹۹۵	۰/۹۳۳	۰/۹۰۶	۰/۹۸۱
	RC	۰/۰۲۶	۰/۰۱۹	۰/۰۷۷	۰/۱۴۰
NoPE	MAE	۰/۸۹۴	۰/۸۹۴	۰/۸۹۴	۰/۸۹۴
	RC	۰/۰۱۸	۰/۰۱۸	۰/۰۱۸	۰/۰۱۸

۳- آزمایش‌ها

۳-۱- مجموعه داده و معیارهای ارزیابی

به‌منظور ارزیابی روش پیشنهادی از مجموعه داده Movielens 1M که به‌صورت آنلاین از وب‌سایت گروپ لنز[۱۶] قابل دریافت است، استفاده شده است. این مجموعه شامل ۱۰۰۰۰۰۰ رتبه است که توسط ۶۰۴۰ کاربر به ۳۹۵۰ فیلم داده شده است. رتبه‌های موجود شامل یک مقیاس عددی ۵ نقطه‌ای است که رتبه ۱ نشان‌دهنده علاقه خیلی کم، رتبه ۲ نشان‌دهنده علاقه کم، رتبه ۳ نشان‌دهنده علاقه متوسط، رتبه ۴ نشان‌دهنده علاقه بالا، و رتبه ۵ نشان‌دهنده علاقه خیلی بالای کاربران است. در این مجموعه داده‌ای، هر کاربر حداقل ۲۰ آیتم رتبه‌داده‌شده دارد.

برای ارزیابی روش پیشنهادی از معیار MAE که با استفاده از رابطه (۵) محاسبه می‌شود، و همچنین، معیار RC که با استفاده از رابطه (۶) محاسبه می‌شود، استفاده شده است.

$$MAE = \frac{\sum_u \sum_i \left| p_{u,i} - r_{u,i} \right|}{n} \qquad (۵)$$

در رابطه فوق $r_{u,i}$ رتبه واقعی کاربر، $p_{u,i}$ رتبه پیش‌بینی‌شده، و n تعداد کل رتبه‌های پیش‌بینی‌شده است.

$$RC = \frac{m}{n} \qquad (۶)$$

m تعداد رتبه‌های پیش‌بینی‌شده در مجموعه تست و n تعداد کل رتبه‌های تست است.

۳-۲- تنظیمات آزمایش

به‌منظور ارزیابی روش پیشنهادی، آن را با روش‌های LMR [۲]، LUC [۲] و همچنین روش پالایش گروهی مبتنی بر کاربر کلاسیک (NoPE) [۲] مقایسه می‌کنیم. در روش پیشنهادی آزمایش‌ها به‌ازای همه کاربران موجود در مجموعه داده‌ای صورت گرفته است. به‌منظور ارزیابی روش پیشنهادی در شرایط شروع سرد کاربران، از روش پیشنهادشده در [۲] استفاده شده است. در مجموعه داده‌ای مورد استفاده برای ارزیابی روش پیشنهادی، هر کاربر حداقل ۲۰ آیتم رتبه داده شده دارد. بنابراین، به‌منظور ارزیابی روش پیشنهادی در شرایط شروع سرد، تعدادی آیتم به‌صورت تصادفی بر اساس روش مطرح‌شده در [۲] برای هر کاربر انتخاب می‌شود. این تعداد آیتم انتخاب شده، کم‌تر از تعداد کل آیتم‌های رتبه‌داده‌شده توسط هر کاربر خواهد بود. برای هرکدام از کاربران تعداد آیتم‌های اولیه پروفایل برابر با مقادیر مختلف N=۲،۳،۴،۵،۱۰ انتخاب می‌شود. تعداد آیتم‌های اضافه شده به پروفایل کاربران I=۵،۱۰،۱۵،۲۰،۵۰ در نظر گرفته می‌شود. در آزمایش‌ها ترکیب آیتم‌های N و I به‌عنوان مجموعه آموزش در نظر گرفته شده است. همچنین، تعداد نزدیک‌ترین همسایگان کاربر هدف یعنی k=۲۵ است. به‌علاوه، مقدار وزن‌های هر یک از داده‌های دموگرافی سن، جنسیت و شغل به ترتیب مقادیر ۰/۳، ۰/۱ و ۰/۶ در نظر گرفته می‌شوند [۱۶]. یکی از پارامترهای مهمی که می‌تواند تأثیر زیادی در کارایی روش پیشنهادی داشته باشد، پارامتر α است. این پارامتر به‌منظور کنترل تأثیر شباهت کسینوسی و شباهت دموگرافی در رابطه (۳) استفاده شده است. در شکل ۲ تأثیر مقادیر مختلف پارامتر α بر روی کارایی روش پیشنهادی بر اساس دو معیار MAE و RC نشان داده شده است. بدین منظور مقادیر مختلف پارامتر α بر اساس مقادیر پارامترهای N=۳، N=۱۰ و I=۱۰ در نظر گرفته شده است. همان‌طور که از این نتایج قابل مشاهده است، مقدار MAE برای هر دو استراتژی GMR و GUC ابتدا کاهش می‌یابد و سپس با افزایش مقدار پارامتر α به‌صورت صعودی افزایش می‌یابد. همچنین، مقدار RC برای هر دو استراتژی GMR و GUC ابتدا افزایش می‌یابد و سپس با افزایش مقدار پارامتر α به‌صورت نزولی کاهش پیدا می‌کند. بر اساس نتایج به‌دست‌آمده از این آزمایش‌ها، مقدار پارامتر α=۰/۵ در نظر گرفته شده است.

۳-۳- نتایج آزمایش‌ها

روش‌های موردنظر در آزمایش‌ها براساس دو معیار MAE و RC با همدیگر مقایسه می‌شوند. لازم به ذکر است، روش پیشنهادی بر اساس نوع استراتژی استفاده‌شده برای گسترش پروفایل به ترتیب، با عناوین GMR و GUC نام‌گذاری شده است.

جدول ۱ نتایج را برای معیارهای MAE و RC با استفاده از N=۳ و مقادیر مختلف I به ازای روش‌های مختلف نشان می‌دهد. همان‌طور که از نتایج جدول ۱ مشخص است، روش پیشنهادی در اکثر موارد بهترین عملکرد را دارد. برای نمونه، روش GMR به‌ازای تمام Iها به‌جز I=۵۰ بهترین دقت را به خود اختصاص داده است. همچنین روش GUC به ازای تمام Iها بهترین نرخ پوشش رتبه را به‌خود اختصاص داده است.

در ادامه این بخش روش‌های مختلف بر اساس دو معیار MAE و RC و به‌ازای N های مختلف و مقدار I=۱۰ با همدیگر مقایسه

می‌شوند. همان‌طور که از نتایج شکل ۳ مشخص است، روش GMR به‌ازای اکثر مقادیر N برای مقدار I=۱۰ بهترین دقت را به خود اختصاص داده است. علاوه‌براین، در شکل ۴ روش GUC به ازای تمام Nها بهترین نرخ پوشش رتبه را به خود اختصاص داده است. نتایج موجود در شکل‌های ۳ و ۴ نشان می‌دهند که روش GUC با این‌که بهترین عملکرد را بر اساس معیار RC در بین سایر روش‌ها داشته است، اما نتوانسته است عملکرد خوبی در مورد معیار MAE داشته باشد. بنابراین، می‌توان نتیجه گرفت که روش GUC به‌دلیل اینکه نرخ پوشش را برای پیش‌بینی‌ها بالا برده است، کارایی آن در مورد معیار MAE کم‌تر شده است. این مسئله در سیستم‌های توصیه‌گر همواره مطرح شده است که معیارهای نرخ پوشش و دقت عکس هم عمل می‌کنند. یعنی، درصورتی‌که یک روش بخواهد نرخ پوشش را بالاتر ببرد، به همان میزان ممکن است از دقت پیش‌بینی‌ها کاسته شود.

شکل‌های ۵ و ۶ به ترتیب مقدار MAE و RC را به‌ازای مقادیر مختلف پارامترهای N و I برای دو روش GMR و GUC نشان می‌دهند. همان‌طور که از شکل ۵ مشخص است، بهترین مقدار MAE برای روش GMR زمانی به‌دست می‌آید که مقدار پارامتر N=۳ و پارامتر I=۱۰ باشد. همچنین بهترین مقدار MAE برای روش GUC زمانی به‌دست می‌آید که مقدار پارامتر N=۱۰ و پارامتر I=۱۵ باشد. علاوه‌براین، شکل ۶ نشان می‌دهد که بهترین مقدار RC برای روش GMR به ازای مقادیر پارامترهای N=۱۰ و I=۱۵ به‌دست می‌آید. برای روش GUC بهترین مقدار RC زمانی به‌دست می‌آید که مقادیر پارامترهای N=۱۰ و I=۱۵ است.

به‌منظور بررسی اهمیت استفاده از اطلاعات دموگرافی کاربران، یک مطالعه آماری بر روی مجموعه داده‌ای انجام شده است. در این مطالعه، دسته‌بندی مختلفی که بر اساس ویژگی‌های جنسیت، سن و شغل افراد در مجموعه داده‌ای وجود دارند، در نظر گرفته شده است. بدین‌منظور، درصد فراوانی کاربرانی که در دسته‌های مختلف قرار می‌گیرند محاسبه شده است. همچنین، میانگین تعداد آیتم‌های مشترک رتبه‌داده‌شده بین کاربران موجود در دسته‌های مختلف محاسبه شده است. نتایج حاصل از این مطالعات آماری در شکل ۷ نشان داده شده است. با توجه به اطلاعات موجود در مجموعه داده‌ای، افراد بر اساس جنسیت در دو گروه مرد و زن قرار می‌گیرند. همان‌طور که از شکل ۷ قابل مشاهده است، میانگین تعداد آیتم‌های مشترک بین افراد با جنسیت مرد بیش‌تر از افراد با جنسیت زن است. به‌طوری‌که، میانگین تعداد آیتم‌های مشترک رتبه‌داده‌شده بین افراد با جنسیت مرد حدوداً برابر ۲۵ است، درصورتی‌که این مقدار برای افراد با جنسیت زن حدوداً ۱۷ است. این نشان می‌دهد که علایق مردان نسبت به زنان بیش‌تر به هم شبیه است. همچنین دسته‌بندی کاربران مختلف بر اساس ویژگی سن، در هفت دسته انجام می‌گیرد که بر اساس تحلیل‌های انجام‌شده، افراد در بازه سنی ۳۴-۲۵ دارای بیش‌ترین درصد فراوانی می‌باشند. علاوه‌براین، میانگین تعداد آیتم‌های مشترک بین کاربران در بازه سنی ۳۴-۲۵ نسبت به سایر بازه‌های سنی بیش‌تر است. درنهایت، تحلیل‌های آماری بر اساس ویژگی شغل افراد انجام گرفته است که کاربران در ۲۱ عنوان شغلی تقسیم‌بندی می‌شوند. نتایج حاصل از این تحلیل نشان می‌دهد که کاربران با عنوان شغلی grad student دارای بیش‌ترین درصد فراوانی در مجموعه داده‌ای می‌باشند. از طرف دیگر، کاربران با عنوان شغلی writer دارای بیش‌ترین مقدار میانگین آیتم‌های مشترک رتبه‌داده‌شده می‌باشند. این تحلیل‌های آماری نشان می‌دهند که استفاده از اطلاعات دموگرافی کاربران در کنار معیارهای شباهت مبتنی بر رتبه کاربران نظیر شباهت کسینوسی، تا چه اندازه می‌تواند در ارائه پیشنهاد به کاربران به‌خصوص کاربران شروع سرد مفید باشد.

۴- نتیجه‌گیری

در این مقاله، یک روش ترکیبی جدید بر اساس ترکیب شباهت دموگرافی و شباهت کسینوسی بین کاربران به‌منظور حل مشکل شروع سرد ارائه شده است. ایده اصلی روش پیشنهادی، گسترش پروفایل کاربران بر اساس استراتژی‌های مختلف به‌منظور ساخت پروفایل‌های با کارایی بالاتر برای کاربران است. نتایج به‌دست‌آمده از آزمایش‌ها نشان‌دهنده عملکرد بهتر روش پیشنهادی در مقایسه با دیگر روش‌ها است. یکی از پیشنهادهایی که برای کارهای آینده می‌توان در نظر گرفت این است که از اطلاعات مربوط به محتوای آیتم‌های مختلف نیز در کنار اطلاعات دموگرافی کاربران استفاده شود. استفاده از این اطلاعات اضافی مربوط به آیتم‌ها، می‌تواند باعث افزایش کارایی سیستم‌های توصیه‌گر به‌خصوص در شرایط شروع سرد کاربران شود.

مراجع

[1] G. Guo, J. Zhang, and D. Thalmann, "Merging trust in collaborative filtering to alleviate data sparsity and cold start," Knowledge-Based Systems, Vol. 57, pp. 57-68, 2014.

[2] V. Formoso, D. Fernandez, F. Cacheda, and V. Carneiro, "Using profile expansion techniques to alleviate the new user problem," Information Processing & Management, Vol. 49, No. 3, pp. 659-672, 2013.

[۳] سیامک عبداله‌زاده، محمدعلی بالافر و لیلی محمدخانلی، «استفاده از خوشه‌بندی و مدل مارکوف جهت پیش‌بینی درخواست آتی کاربر در وب»، مجله مهندسی برق دانشگاه تبریز، دوره ۴۵، شماره ۳، صفحه ۸۹-۹۶، تبریز، ۱۳۹۴.

[4] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," Advances in Artificial Intelligence, Vol. 2009, No. 4, pp. 1-19, 2009.

[5] J. Bobadilla, A. Hernando, F. Ortega, and J. Bernal, "A framework for collaborative filtering recommender systems," Expert Systems with Applications, Vol. 38, No. 12, pp. 14609-14623, 2011.

[6] S. Ghazarian and M. A. Nematbakhsh, "Enhancing memory-based collaborative filtering for group recommender systems," Expert Systems with Applications, Vol. 42, No. 7, pp. 3801-3812, 2015.

[7] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: an open architecture for collaborative filtering of netnews," in Proceedings of the 1994 ACM conference on Computer supported cooperative work,

Chapel Hill, North Carolina, USA, ACM, pp. 175-186, 1994.

[8] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, Madison, Wisconsin, USA, ACM, pp. 43-52, 1998.

[9] M. Swaby, P. Dew, and P. Kearney, "Model-based Construction of collaborative systems," BT technology journal, Vol. 17, No. 4, pp.78-90, 1999.

[10] C. Birtolo and D. Ronca, "Advances in clustering collaborative filtering by means of fuzzy c-means and trust," Expert Systems with Applications, Vol. 40, No. 17, pp. 6997-7009, 2013.

[11] M. H. Park, J. H. Hong, and S. B. Cho, "Location-based recommendation system using bayesian user's preference model in mobile devices," in Proceedings of the 4th international conference on Ubiquitous Intelligence and Computing (UIC'07), Hong Kong, China, Springer, pp. 1130-1139, 2007.

[12] L. Gao and C. Li, "Hybrid personalized recommended model based on genetic algorithm," in 4th International Conference on Wireless Communications, Networking and Mobile Computing, Dalian University of Technology and Wuhan University, China, IEEE, pp. 1-4, 2008.

[13] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutierrez, "Recommender systems survey," Knowledge-Based Systems, Vol. 46, pp. 109-132, 2013.

[14] L. H. Son, "Dealing with the new user cold-start problem in recommender systems: A comparative review," Information Systems, Vol. 58, pp. 87-104, 2016.

[۱۵] مصطفی رجب‌زاده و رضا رافع، «ارائه یک سیستم توصیه‌گر ترکیبی برای تجارت الکترونیک»، مجله مهندسی برق دانشگاه تبریز، دوره ۴۵، شماره ۴، صفحه ۸۵-۹۱، تبریز، ۱۳۹۴.

[16] B. Lika, K. Kolomvatsos, and S. Hadjiefthymiades, "Facing the cold start problem in recommender systems," Expert Systems with Applications, Vol. 41, No. 4, pp. 2065-2073, 2014.

[17] K. Ji and H. Shen, "Jointly modeling content, social network and ratings for explainable and cold-start recommendation," Neurocomputing, Vol. 218, pp. 1-12, 2016.

[18] M. Y. H. Al-Shamri, "User profiling approaches for demographic recommender systems," Knowledge-Based Systems, Vol. 100, pp. 175-187, 2016.

[19] L. Safoury and A. Salah, "Exploiting user demographic attributes for solving cold-start problem in recommender system," Lecture Notes on Software Engineering, Vol. 1, No. 3, pp. 303-307, 2013.

[20] Y. Wang, S. C. F. Chan, and G. Ngai, "Applicability of demographic recommender system to tourist attractions: A case study on trip advisor," IEEE Computer Society, Vol. 3, pp. 97-101, 2012.

[21] D. Anand, and K. K. Bharadwaj, "Utilizing various sparsity measures for enhancing accuracy of collaborative recommender systems based on local and global similarities," Expert systems with applications, Vol. 38, No. 5, pp. 5101-5109, 2011.

[22] H. J. Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," Information Sciences, Vol. 178, No. 1, pp. 37-51, 2008.

زیرنویس‌ها

۱ Recommender system

۲ Information filtering system

۳ Personalization

۴ Overall sparsity measure

۵ User-specific sparsity measure

۶ User-item specific sparsity measures

۷ Unified measure of sparsity

۸ Proximity–Impact–Popularity

۹ Item-global

۱۰ Item-local

۱۱ User-local

۱۲ Local most-rated

۱۳ Local user-local clustering

۱۴ Global Most Rated

۱۵ Global User Clustering

۱۶ http://grouplens.org/datasets/movielens/1m/