فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)
انجمن علمی تجارت الکترونیکی ایران
واحد زنجان

ارائه روشی جدید برای کشف نزدیکترین همسایگی در سیستم‌های توصیه‌گر مبتنی بر فیلترینگ مشارکتی

مقاله پژوهشی

DOR: 20.1001.1.27832570.1400.2.1.3.4

احسان‌اله کوزه‌گر[*1]، هادی یارمحمدی[2]، مهدی بازرگانی[3] ، زینب همایونپور[4]،

۱ - گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی شرق گیلان، دانشگاه گیلان، گیلان، ایران آدرس ایمیل:.Kozegar@guilan.ac.ir

۲ - دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، سمنان، ایران

۳، ۴- گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد زنجان، زنجان، ایران

چکیده: سیستم‌های توصیه‌گر با تحلیل و بررسی داده‌های متعلق به کاربران، یکسری آیتم‌های خاص را برمبنای علایق به کاربران پیشنهاد می‌کنند. هدف از آنالیز داده‌های مربوط به کاربران، استخراج الگوهای هر کاربر به منظور پیش‌بینی آیتم‌ها می‌باشد. یکی از مهمترین روش‌ها در سیستم‌های توصیه‌گر، روش فیلترینگ مشارکتی است. در سیستم‌های توصیه‌گر مبتنی بر فیلترینگ مشارکتی از معیارهای شباهت جهت کشف کردن کاربران مشابه با کاربر جدید برای ارائه پیشنهاد استفاده می‌شود. از چالش‌های سیستم‌های توصیه‌گر مبتنی بر فیلترینگ مشارکتی می‌توان به فاکتورهای شباهت و تشخیص همسایگی اشاره کرد. در این مقاله از روش نزدیک‌ترین همسایه به منظور تشخیص همسایگان مشابه به کاربر جدید برمبنای فاصله استفاده می‌کنیم. مدل پیشنهادی که برگرفته از روش کاربر-آیتم است، امتیاز قلم دادهبرمبنای فاصله محاسبه می شود و نزدیکترین فاصله به منظور تشابه انتخاب می‌شود. در مدل پیشنهادی، تشخیص کاربران مشابه براساس ماتریس کاربر-آیتم توسط فاصله اقلیدسی انجام می‌شود. آزمایشات مدل پیشنهادی برروی مجموعه داده Movielens که شامل ۱۶۸۲ آیتم است انجام شده است. برای ارزیابی از معیارهای دقت، فراخوانی، F1، میانگین خطای مطلق و میانگین خطای مربعات ریشه استفاده شده است. میانگین خطای مطلق در مدل پیشنهادی در مقایسه با شباهت پیرسون و کسینوسی کمتر است و مقدار آن برابر با ۰/۷۳۱۵ می‌باشد. در نتیجه دقت مدل پیشنهادی در تشخیص تشابه و پیش‌بینی بیشتر است.

واژه‌های کلیدی: سیستم‌های توصیه‌گر، فیلترینگ مشارکتی، نزدیکترین همسایه، میانگین خطای مطلق

Providing A New Method to Discover the Closest Neighborhood in Recommendation Systems Based on Collaborative Filtering

Ehsan Kozegar[1], Hadi Yar Mohammadi[2], Mahdi Bazergani[3], Zeinab Homaounpour[3]

1- Department of Computer Engineering, Faculty of Technology and Engineering, University of Guilan, Guilan, Iran.
2- Faculty of Computer Engineering, Shahrood University of Technology, Semnan, Iran
3, 4- Faculty of Electrical and Computer Engineering, Islamic Azad University of Zanjan, Zanjan

Abstract: Recommender systems, by analyzing data belonging to users, suggest a number of specific items based on interests to users. The purpose of analyzing user data is to extract each user's patterns in order to predict items. One of the most important methods in recommender systems is participatory filtering method. In collaborative filtering recommendation systems, similarity criteria are used to identify users similar to the new user to submit a proposal. One of the challenges of collaborative filtering-based recommendation systems is similarity and neighborhood detection factors. In this paper, we use the nearest neighbor method to identify similar neighbors to the new user based on distance. The proposed model, which is derived from the user-item method, calculates the data font score based on the distance, and the closest distance is selected for similarity. In the proposed model, the identification of similar users based on the user-item matrix is done by the Euclidean distance. The experiments of the proposed model were performed on the Movielens dataset which contains 1682 items. Accuracy, recall, F1, mean absolute error and mean root square error were used for evaluation. The average absolute error in the proposed model is less than Pearson and cosine similarity and its value is equal to 0.7315. As a result, the proposed model is more accurate in detecting similarity and prediction.

Keywords: Recommender Systems, Collaborative Filtering, nearest neighbor, Mean Absolute Error

تاریخ ارسال مقاله: ۱۳۹۹/۰۸/۱۸

تاریخ پذیرش مقاله: ۱۳۹۹/۰۹/۲۷

*: نویسنده مسئول

۱. مقدمه

سیستم‌های توصیه‌گر یک مدل خاص از سیستم‌های پشتیبان تصمیم (سیستم‌های اطلاعاتی) می‌باشند که نقش مهمی در راهنمایی و هدایت کاربران در میان حجم عظیمی از انتخاب‌های ممکن دارند تا محصولات مناسب را برطبق اولویت‌ها و علایق خود انتخاب کنند [۱]. سیستم‌های توصیه‌گر در حوزه‌های مختلفی جهت شخصی‌سازی پیشنهادها بکار گرفته شده‌اند که از جمله آن‌ها می‌توان به کتاب، موزیک، فیلم، اخبار، مقالات و غیره اشاره کرد. سیستم‌های توصیه‌گر می‌توانند به عنوان برنامه‌هایی تعریف شوند که مناسب‌ترین قلم داده(محصول یا سرویس) را به کاربران توصیه کنند [۲]. این توصیه‌ها با پیش‌بینی سلایق کاربر به یک نمونه خاص براساس اطلاعات مرتبط در مورد نمونه‌ها، کاربران و ارتباطات بین کاربران و نمونه‌ها ارائه می‌شود. سیستم‌های توصیه‌گر با توجه به سلایق شخصی افراد با در نظر گرفتن انتخاب‌های قبلی وی خدماتی را به آن‌ها ارائه می‌نمایند. هدف از توسعه سیستم‌های توصیه‌گر، کاهش سربار اطلاعات به وسیله بازیابی نزدیکترین اطلاعات و خدمات از حجم انبوهی از داده‌ها و در نتیجه ارائه خدمات شخصی‌شده می‌باشد. کاربرد همه منظوره از وب باعث شده تا کاربران با حجم عظیمی از داده‌ها و اطلاعات سر و کار داشته باشند. کاربران در مواجهه با این حجم اطلاعات دچار سردرگمی در انتخاب اطلاعات مورد نظر خود می‌شوند. سیستم‌های توصیه‌گر ابزار بسیار مهم و کارآمدی در وب سایت‌های تجارت الکترونیک محسوب می‌شوند و به کاربران در کشف قلم دادهمورد علاقه‌شان کمک می‌کنند. هدف اصلی سیستم‌های توصیه‌گر تولید پیش‌بینی‌های دقیق می‌باشد. این سیستم‌ها می‌توانند پیشنهادهایی را براساس پروفایل کاربران و یا اولویت‌های قبلی آن‌ها ارائه کنند و یا اینکه بر انتخاب‌های سایر افراد به عنوان داور تکیه کنند. مهم‌ترین ویژگی یک سیستم توصیه‌گر قابلیت پیش‌بینی اولویت‌ها و سلیقه‌های یک کاربر به وسیله تحلیل رفتار آن و رفتار سایر کاربرها جهت تولید پیشنهادهای شخصی شده است [۳].

فیلترینگ مشارکتی معمول‌ترین روش مورد استفاده در سیستم‌های توصیه‌گر است. این روش مبتنی بر ارزیابی کاربرانی است که علاقه‌های مشابهی دارند و ایده اصلی این سیستم‌ها این است که کاربرانی که آیتم‌های مشابهی را در گذشته انتخاب کرده‌اند احتمالا ارجحیت‌های مشابهی دارند. سیستم‌های فیلترینگ مشارکتی از داده‌های قدیمی مرتبط با اولویت‌های کاربر یا رفتار آن استفاده می‌کنند تا بتوانند رفتار کاربران جدید را پیش‌بینی کنند. هدف از فیلترینگ مشارکتی پیش‌بینی رتبه مربوط به اقلام‌های رتبه‌بندی نشده در ماتریس کاربر-آیتم می‌باشد [۴]. در فیلترینگ مشارکتی فرض براین است که اگر دو کاربر رتبه مشابهی را به یک آیتم (کتاب، فیلم، موزیک و...) بدهند یا رفتارهای مشابهی (خرید کردن، مطالعه کردن، تماشا کردن و غیره) را داشته باشند، آنها می‌توانند اقلام‌های مشابه را رتبه‌بندی کنند. مبتنی بر این فرض، فیلترینگ مشارکتی برمبنای نرخ تشابه کاربران و اقلام‌ها به پیش‌بینی رتبه اقلام‌های رتبه‌بندی نشده می‌پردازد. در این سیستم‌ها یک مدل یا سند برمبنای ارزیابی مجموعه‌ای از کاربران ایجاد می‌شود در واقع کاربران را گروه‌بندی می‌کند، هر گروه شامل کاربرها و اقلام‌هایی است که شبیه به هم می‌باشند. فیلترینگ مشارکتی دارای ضعف‌هایی زیر است: در این روش سیستم باید پیشنهاداتی را برمبنای رتبه‌ی کاربر به سایر اقلام‌ها به وی ارائه دهد. از معایب دیگر این روش می‌توان به اضافه شدن قلم داده جدید به مجموعه‌ی اقلام‌ها اشاره کرد. تا زمانیکه یک قلم داده دارای رتبه‌ای نباشد، آن قلم داده به کاربر جدید پیشنهاد نمی‌گردد. به علاوه مشکل پراکندگی در این الگوریتم وجود دارد. یعنی تعداد رتبه‌هایی که از گذشته بدست آمده بسیار کمتر از تعداد مورد نیاز برای پیشگویی است [۵].

فیلترینگ مشارکتی به طور وسیعی نسبت به روش‌های دیگر در سیستم‌های توصیه‌گر مورد استفاده قرار گرفته است. این سیستم‌ها پیش‌بینی را از طریق سابقه همسایگان (کاربران یا آیتم‌ها) محاسبه می‌کنند و انتخاب صحیح نزدیک‌ترین همسایه که بیشترین شباهت را به کاربر هدف داشته باشند تاثیر زیادی بر دقت پیش‌بینی‌ها خواهد داشت. در روش پیشنهادی، الگوریتم نزدیکترین همسایه برمبنای تکنیک‌های تشخیص فاصله در جهت کسب نتایج بهتر استفاده می‌شود.

در سیستم‌های توصیه‌گر بیشتر از روش فیلترینگ مشارکتی استفاده شده است. مشکلات مهم این سیستم‌ها، پراکندگی، مقیاس‌پذیری و هم‌معنایی می‌باشد که در دقت و کیفیت پیشنهاددهی تاثیر منفی می‌گذارد. با توجه به این که افزایش دقت پیش‌بینی در سیستم‌های توصیه‌گر فیلترینگ مشارکتی نقش اساسی در بسیاری از زمینه‌های زندگی از جمله تجارت، پزشکی و... ایفا می‌کند، تاکنون روش‌ها و الگوریتم‌های متعددی جهت رفع این مشکلات ارائه شده است، اما هنوز چالش مهم این سیستم‌ها، افزایش دقت آن می‌باشد. بنابراین، نیاز به ارائه‌ی روش‌های هوشمند جهت بهبود دقت این سیستم‌ها لازم و اساسی می‌باشد و در این تحقیق در راستای رفع این نیاز از روش مبتنی بر مدل استفاده می‌شود. الگوریتم‌های مبتنی بر مدل، پیش‌بینی رتبه‌بندی را با استفاده از روش‌های یادگیری ماشین و آماری، برای یادگیری یک مدل از داده‌ها انجام می‌دهند. جنبه جدید بودن این تحقیق در استفاده از الگوریتم نزدیکترین همسایه و بکارگیری روش‌های مختلف فاصله برای تشخیص همسایگان می‌باشد.

۲. مطالعات پیشین

در طرح [۶] پروین و همکارانش از الگوریتم بهینه‌سازی کلونی مورچه به منظور پیش‌بینی رأی‌های از دست رفته مبتنی بر فیلترینگ مشارکتی استفاده کرده‌اند. روش پیشنهادی شامل سه مرحله اصلی است: در مرحله اول، کاربران با در نظر گرفتن مقدار ارزش و روابط اعتماد رتبه‌بندی می‌شوند. سپس، در مرحله دوم، الگوریتم بهینه‌سازی کلونی مورچه برای تعیین مقادیر مناسب وزن به کاربران مورد استفاده قرار می‌گیرد. یک مجموعه از کاربران مشابه در مرحله سوم فیلتر می‌شوند تا در پیش‌بینی رتبه‌های ناشناخته برای کاربر هدف مورد استفاده قرار گیرند. به عبارت

دیگر، برای سرعت بخشیدن به شناسایی کاربران مشابه، روش پیشنهادی ابتدا بخش اکثریت کاربران غیرمشابه را فیلتر می‌کند و سپس الگوریتم بهینه‌سازی کلونی مورچه را فقط در یک مجموعه کاهش یافته از کاربران اجرا می‌کند تا آنها را وزن‌دهی کند. در هر تکرار موچه‌ها ارزیابی می‌شوند و برمبنای انتخاب‌شان، بهترین مقدار برای موارد ناشناخته لحاظ می‌گردد. شبیه‌سازی برروی سه مجموعه داده Film-Trust، E-pinions و Ciao انجام شده است. نتایج نشان داده است که الگوریتم بهینه‌سازی کلونی مورچه از مقدار خطای کمتری در مقایسه با تکنیک‌های دیگر بهره‌مند است.

در طرح [۷] کانت و همکارانش از الگوریتم خوشه‌بندی k-means به منظور گروه‌بندی آیتم‌های مشابه استفاده کرده‌اند که هدف آنها یافتن مراکز خوشه بهینه به منظور گروه‌بندی آیتم‌ها در گروه‌های مشابه است. الگوریتم خوشه‌بندی k-means برروی ماتریس داده‌های کاربر مورد استفاده قرار می‌گیرد تا محدوده مورد نظر را تعیین کند. الگوریتم k-means خوشه‌بندی قلم داده را براساس اندازه‌گیری فاصله تعریف شده بین دو آیتم مختلف پیدا می‌کند. الگوریتم خوشه‌بندی k-means گروه‌هایی از قلم داده را شناسایی می‌کند به طوریکه فقط متعلق به یک خوشه باشند. شبیه‌سازی برروی مجموعه داده MovieLens100k انجام شده است. نتایج نشان می‌دهد که الگوریتم k-means بهبود داده شده در مقایسه با k-means معمولی خطای کمتری دارد.

در طرح [۸] واسید و همکارش یک مدل برمبنای الگوریتم بهینه‌سازی اجتماع ذرات و فازی برای فیلترینگ مشارکتی پیشنهاد داده‌اند. در طرح آنها از الگوریتم بهینه‌سازی اجتماع ذرات برای وزندهی به ویژگی‌ها و از مجموعه فازی برای انتخاب ویژگی‌های موثر استفاده شده است. در این کار، با استفاده از الگوریتم بهینه‌سازی ذرات، اولویت‌ها برمبنای رتبه ایجاد می‌شوند. پس از پیدا کردن وزن مناسب برای ویژگی رتبه، بهترین تشابه به کاربر فعال نسب داده می‌شود. شبیه‌سازی برروی مجموعه داده MovieLens انجام شده است. نتایج نشان داده که روش پیشنهادی در مقایسه با روش‌های دیگر از درصد خطای کمتری بهره‌مند است.

در طرح [۹] جو و همکارش از الگوریتم زنبور مصنوعی برای خوشه‌بندی K-means استفاده کرده‌اند. در فرایند خوشه‌بندی از الگوریتم زنبور مصنوعی برای رفع مشکل ناشی از خوشه‌بندی K-means استفاده شده است. سپس شباهت بین کاربران یک خوشه، توسط شباهت کسینوسی بهبود داده شده، محاسبه گردیده است. شبیه‌سازی مدل ترکیبی برروی مجموعه داده MovieLens انجام شده است.

در طرح [۱۰] آر و بوستانسی به منظور کاهش خطای پیش‌بینی از الگوریتم ژنتیک و معیار شباهت بردار کسینوسی استفاده کرده‌اند. از الگوریتم ژنتیک برای کاربران همسایه بسیار مشابه وزندهی شده و با استفاده از معیار شباهت و وزن‌های به دست آمده، دقت در انتخاب همسایگان افزایش یافته است و با استفاده از جهش ژنتیکی به کاربران جدید امکان اضافه شدن به لیست کاربران مشابه را داده و به این ترتیب باعث بهبود مشکل شروع سرد شده است. شبیه‌سازی مدل ترکیبی برروی مجموعه داده MovieLens انجام شده است. نتایج نشان داده که مدل پیشنهادی در مقایسه با مدل‌های سنتی از درصد خطای کمتر و افزایش دقت برخوردار بوده است.

در طرح [۱۱] کوهی و کیانی از روش خوشه‌بندی c میانگین فازی یا FCM برای فیلترینگ مشارکتی مبتنی برکاربر استفاده کرده‌اند و عمکرد آن در برابر روش‌های خوشه‌بندی مختلف موردی بررسی قرار گرفته است. رویه این مدل به پنج بخش تقسیم می‌شود: ۱) آماده سازی داده‌ها: خواندن داده و پاکسازی داده‌ها. ۲) خوشه‌بندی کاربر: الگوریتم خوشه‌بندی FCM برای تولید P بخش از کاربران، اعمال شده است. این خوشه‌بندی به عنوان پایه‌ای برای پیدا کردن خوشه‌های دیگر، استفاده شده است. ۳) خوشه‌بندی پویای فازی: خوشه‌های فازی در قالب‌های زمانی مختلف، برای یک کاربر فعال، پیدا شده و درجه پویایی عضویت، محاسبه شده است. ۴) انتخاب همسایه: همسایگی برای یک کاربر داده شده براساس خوشه‌بندی فازی، تعیین می‌شود. ۵) توصیه: رتبه‌بندی کاربر فعال از آیتم رتبه‌بندی نشده براساس رتبه‌بندی‌های همسایه‌ها، پیش‌بینی می‌شود. داده‌های ورودی در این روش، از یک ماتریس کاربر به رتبه‌بندی، وارد شده‌اند. مجموعه داده مووی لنز(MovieLens) برای مقایسه الگوریتم‌های خوشه‌بندی مختلف مورد استفاده قرار گرفته است. و روش‌ها از نظر دقت و صحت و بازخوانی ارزیابی شده‌اند. نتایج شبیه‌سازی نشان داده که ترکیبی از روش خوشه‌بندی فازی با دیفازی کردن مرکز ثقل و ضریب همبستگی پیرسون منجر به توصیه‌های بهتری به نسبت سایر روش‌ها شده است.

در طرح [۱۲] مدل خوشه‌بندی بهبودیافته-میانگین به منظور کاهش مشکل پیچیدگی محاسباتی پیشنهاد شده است که کاربران را به خوشه‌هایی همگن و منسجم تقسیم‌بندی کرده است. خوشه‌بندی برمبنای معیار شباهت فاصله انجام شده است که کاربران درون هر خوشه دارای شباهت بالا و شباهت بین خوشه‌ها در سطح پایینی می‌باشد. مدل پیشنهادی به صورت بازگشتی داده‌های حجیم نظرات را به دو زیرخوشه تقسیم کرده است. درخت حاصل از این بخش‌بندی، یک درخت دودویی غیرمتعادل بوده است که هر برگ آن دارای یک ماتریس شباهت و در هر گره داخلی، مراکز نظرات زیردرخت‌ها ثبت شده است. به منظور پیش‌بینی نظر، ابتدا خوشه‌ای که کاربر به آن تعلق داشته است مشخص شده و سپس از نظرات موجود در همان خوشه برای تشخیص نظر نزدیک به کاربر فعال استفاده شده است. کارایی مدل پیشنهادی به تعداد خوشه‌ها و اندازه خوشه‌ها وابسته بوده است. در زمان اجرا، هر چقدر اندازه خوشه‌ها کوچک‌تر باشد، بار محاسباتی کمتر ولی از طرفی دیگر کیفیت پیش‌بینی نظرات کاهش پیدا کرده است.

در [۱۳] از الگوریتم خفاش برای محاسبه وزن آیتم‌ها (ویژگی‌ها) استفاده شده است تا همسایگی بهتری برای کاربر فعال پیدا شود. این تکنیک برای وزندهی به آیتم‌ها با استفاده از الگوریتم خفاش در دستیابی به

توصیه‌های بهتر کمک کرده است. عملکرد این سیستم با عملکرد سیستم مبتنی بر کلونی زنبورهای مصنوعی نیز مقایسه شده است. نتایج نشان داد که الگوریتم خفاش از نظر میانگین خطای مطلق و معیار F1 در مقایسه با الگوریتم کلونی زنبور مصنوعی در حدود ۶,۹ درصد بهتر بوده است.

در [۱۴]، یک مدل فیلترینگ مشارکتی مبتنی بر یادگیری عمیق ارائه شده است. این مدل شامل دو بخش است: بخش اول این مدل مربوط به ویژگی‌های کاربران و آیتم‌ها است و از آنها به عنوان ورودی در رتبه‌بندی شبکه عصبی عمیق استفاده می‌شود، که رتبه‌بندی‌ها را پیش‌بینی می‌کند. این رتبه‌بندی‌ها ورودی بخش دوم را تشکیل می‌دهند که یک رتبه‌بندی مبتنی بر شبکه عصبی عمیق است و برای پیش‌بینی رتبه‌بندی کلی استفاده می‌شود. آزمایشات بر روی یک مجموعه داده واقعی نشان داده است که مدل پیشنهادی از روش‌های دیگر بهتر عمل کرده است و مقدار خطای کمتری داشته است.

در [۱۵] یک سیستم توصیف فیلم مبتنی بر مدل ترکیبی که از k-means و ژنتیک تشکیل شده است برای پارتیشن‌بندی فضای کاربران ارائه شده است. این روش از تجزیه و تحلیل مؤلفه‌های اصلی داده برای کاهش تراکم فضای جمعیت فیلم استفاده می‌کند که می‌تواند پیچیدگی محاسبات را کاهش دهد. الگوریتم K-Means ابتدا K عضو(که K تعداد خوشه‌ها است) را به صورت تصادفی از میان N عضو انتخاب می‌نماید و آنها را به عنوان مراکز خوشه‌ها در نظر می‌گیرد. سپس N-K عضو باقیمانده به نزدیکترین خوشه تخصیص می‌یابند. بعد از تخصیص همه اعضا مجدداً مراکز خوشه‌ها محاسبه شده و اعضا با توجه به میزان نزدیکی (شباهت) به یکی از خوشه‌ها تخصیص می‌یابند و این کار تا زمانی که مراکز خوشه‌ها ثابت بمانند، ادامه می‌یابد. با تکرار همین روال می‌توان در هر تکرار با میانگین‌گیری از داده‌ها مراکز جدیدی برای آنها محاسبه کرد و مجدداً داده‌ها را به خوشه‌های جدید نسبت داد. ارزش رتبه‌بندی تخمین زده شده برای یک فیلم بدون رتبه‌بندی توسط کاربر Ua طبق معادله(۱) تعریف شده است.

$$P_{Ua,item} = \bar{R}_u + \frac{\sum_{y \in C_x} sim(U_a, y) \times (R_{y,i} - \bar{R}_y)}{\sum_{y \in C_x} (|sim(U_a, y)|)} \quad (۱)$$

در معادله(۱) میانگین رتبه‌بندی برمبنای Ua، Cx مجموعه‌ای از همسایه‌ها متعلق به خوشه Ua، پارامتر $\bar{R}_y$ میانگین رتبه‌بندی برمبنای امتیازهای دریافت شده از همسایگان Ua، پارامتر $sim(U_a, y)$ تابع تشابه برمبنای معیار پیرسون برای درجه تشابه دو کاربر. نتایج آزمایش بر روی مجموعه داده Movielens نشان داده است که روش پیشنهادی تواسته از لحاظ دقت، عملکرد بالایی را ارائه دهد و در مقایسه با روش‌های موجود، پیشنهادهای قابل اطمینان‌تر و شخصی‌تری تولید کند.

در [۱۶] از یک روش فیلترینگ گروهی برای ارائه توصیه به کاربران استفاده شده است. در اکثر روش‌های فیلترینگ گروهی از متدهایی مانند ضریب پیرسون و روش‌های مبتنی بر کسینوس برای بدست آوردن میزان شباهت دو کاربر استفاده می‌شود. اما در این تحقیق فاصله کاربران با استفاده از هفت روش مختلف (فاصله اقلیدسی، فاصله چبیشف، فاصله گاور، فاصله سورنسن، فاصله کانبرا، فاصله لورنتزین و فاصله سیتی بلاک) اندازه گیری و محاسبه شده است. با توجه به فاصله محاسبه شده از دو الگوریتم k نزدیک‌ترین همسایه و الگوریتم k-means برای مشخص شدن کاربران مشابه به کاربر هدف استفاده شده است. نتایج نشان داده است که دقت این مدل تقریبا ۶۰ درصد است و میانگین خطا چیزی در حدودی ۲۵ درصد بوده است.

در [۱۷] یک سیستم پیشنهادی برای فیلترینگ مشارکتی برمبنای ترکیب میانگین فازی و الگوریتم جلبک مصنوعی پیشنهاد شده است. برای یافتن شباهت بین دو کاربر از ضریب همبستگی پیرسون چندسطحی پیشرفته استفاده شده است. علاوه بر این، رتبه به فیلم‌هایی که هنوز کاربر به احتمال زیاد رتبه‌ای برای آنها اختصاص نداده است پیدا می‌شود. سیستم پیشنهادی در مقایسه با روش‌های دیگر، موفق به ارائه توصیه‌هایی با کیفیت و دقت بهتر بوده است. سیستم پیشنهادی در چهار مجموعه داده واقعی، آزمایش و ارزیابی شده است: Movielens 100، 000، Movielens یک میلیون، Jester و Epinion. سیستم پیشنهادی برای هر چهار مجموعه داده توصیه‌های بهتری ارائه داده است. با استفاده از معیارهای ارزیابی از جمله میانگین خطای مطلق (MAE)، دقت و فراخوان، کارایی سیستم برآورد شده است.

در [۱۸] یک سیستم توصیه‌گیر مبتنی بر الگوریتم ژنتیک[1] ارائه شده است که به اطلاعات معنایی و داده‌های رتبه‌بندی قبلی بستگی دارد. ایده اصلی این تحقیق در ارزیابی لیست توصیه‌های احتمالی به جای ارزیابی آیتم‌ها و سپس تشکیل لیست توصیه‌ها است. BLIGA از الگوریتم ژنتیک برای یافتن بهترین لیست موارد به کاربر فعال استفاده می‌کند. بنابراین، هر کروموزم لیست پیشنهادی کاندیداها را نشان می‌دهد. BLIGA سلسله مراتبی، کروموزم‌ها را با استفاده از سه تابع برازندگی ارزیابی کرده است. تابع اول از اطلاعات معنایی آیتم‌ها استفاده کرده تا قدرت شباهت معنایی بین آیتم‌ها را تخمین بزند. تابع دوم شباهت سطح رضایت کاربران را تخمین زده است. تابع سوم به رتبه‌بندی پیش‌بینی شده بستگی داشته است تا بهترین لیست توصیه‌ها را انتخاب نماید. نتایج نشان داده است که برتری BLIGA و توانایی آن برای دستیابی به پیش‌بینی دقیق‌تر از روش‌های دیگر بوده است.

در [۱۹] یک چارچوب جدید برمبنای فازی به منظور مسئله شروع سرد پیشنهاد شده است. در این روش برای پیشنهاد آیتم به کاربران جدید از قوانین فازی استفاده شده است. در معادله(۲) p و q مجموعه اعداد فازی هستند که برمبنای آیتم‌های قبلی تنظیم شده‌اند.

$$S_{u_{j,u_k}}(i_p) = 1 - \frac{1}{\sqrt{5}}\left(\sum_{h=1}^{5}(q_{jp}^{h} - q_{kp}^{h})^2\right)^{1/2} \quad (۲)$$

مدل فازی شامل مراحل استنتاج فازی، مراحل تصمیم‌گیری و توصیه می‌باشد. نتایج نشان داده است که مدل پیشنهادی میانگین خطای پایین و درصد دقت بالایی بالایی داشته است.

در [۲۰] به منظور افزایش دقت پیش‌بینی بهبود سیستم‌های فیلترینگ مشارکتی، از طرحی مبتنی بر الگوریتم ژنتیک استفاده شده است. در این طرح به طور انتخابی مقادیر اندازه‌گیری شباهت با بالاترین مقدار انتخاب می‌شوند و پروفایل کاربری در داخل یک کروموزوم، کد می‌شود. مدل به سه فاز تقسیم می‌شود: فاز اول: فاز جمع‌آوری اطلاعات کاربری است که در آن برای یک فرد جدید، سیستم از او می‌خواهد که به منظور جمع‌آوری ویژگی‌ها شخصی و همچنین رتبه‌بندی‌ها از یک سری از فیلم‌ها، ثبت نام کند. فاز دوم ایجاد ویژگی‌های پروفایل کاربر است که برای آن دسته از فیلم‌هایی که کاربر رتبه بندی کرده است. فاز سوم که فاز توصیه‌گر الگوریتم ژنتیک نامیده شده است، دارای توابع مرتبط با الگوریتم ژنتیک است که توصیه مناسب را از طریق انتخاب و وزندهی ویژگی‌ها، دنبال می‌کند و سپس به انجام توصیه می‌پردازد. در این روش برای اندازه‌گیری شباهت نیز از ضریب همبستگی پیرسون استفاده شده است. برای ارزیابی این روش، آزمایش‌هایی بر روی مجموعه داده MovieLens برمبنای معیار دقت انجام شده است و مشخص شده است که این روش با مقدار دقت بین ۸۱/۲۸ تا ۸۱/۷۷ درصد نسبت به روش پیرسون پایه با مقدار دقت ۶۹/۲ درصد، دقت بالاتری را نشان داده است. مزیت این روش این است که همیشه یک جواب نسبتا خوب تولید می‌کند و دقت نسبت به روش‌های پیشین افزایش یافته است. عیب این روش این است که نیاز به حافظه و زمان زیادی دارد و لزوما بهینه‌ترین جواب را به دست نمی‌آورد.

۳. مدل پیشنهادی

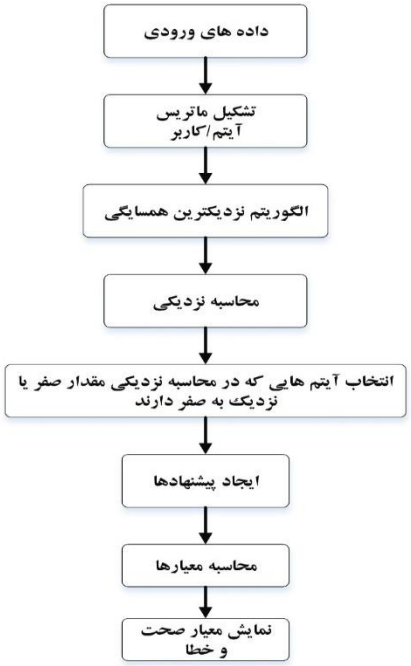
فیلترینگ مشارکتی، یکی از محبوب‌ترین و کارآمدترین مدل‌ها برای تعریف پیشنهاد بالقوه به کاربران جدید است. دو نوع الگوریتم فیلترینگ مشارکتی وجود دارد: سیستم توصیه‌گر مبتنی بر کاربر، که آیتم‌ها را با رتبه‌بندی شباهت آیتم‌های مشترک بین کاربران قبلی به کاربران جدید پیشنهاد می‌کند. فیلترینگ مشارکتی مبتنی بر آیتم، با در نظر گرفتن رتبه آیتم‌ها و تاریخچه آنها، آیتم را به کاربران جدید توصیه می‌کند. اگرچه فیلترینگ مشارکتی با موفقیت در بسیاری از زمینه‌های تجاری مورد استفاده قرار گرفته‌اند، اما چندین اشکال اصلی فیلترینگ مشارکتی، به ویژه کمبود داده‌های رتبه‌بندی، چالش جدی برای صحت و جهانی بودن این روش‌ها ایجاد کرده است. به طور خاص، بیشترین تعداد رتبه‌بندی برای هر کاربر در بسیاری از آیتم‌ها وجود ندارد و عملکرد فیلترینگ مشارکتی به همراه افزایش تعداد آیتم بدون رتبه در مجموعه داده‌های آموزشی، کاهش می‌یابد.

الگوریتم نزدیک‌ترین همسایه در بیشتر موارد به منظور دسته‌بندی آیتم‌های مشابه به کار می‌رود، هرچند که می‌توان از آن برای تخمین و پیش‌بینی نیز استفاده نمود. در الگوریتم نزدیک‌ترین همسایه، آیتم‌ها برمبنای فاصله کشف می‌شوند. معیار فاصله برای تعیین تشابه بین آیتم‌ها استفاده می‌شود. اگر دو بردار q و p داشته باشیم از فاصله اقلیدسی طبق معادله(۳) برای بدست آوردن فاصله بین دو آیتم q_i و p_i استفاده می‌شود.

$$d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (۳)$$

در الگوریتم نزدیک‌ترین همسایه، باید در بین آیتم‌ها، آیتمی انتخاب شود که برای آیتم بدون امتیاز کمترین فاصله را داشته باشد. درنتیجه، پس‌ازآنکه فاصله اقلیدسی بین آیتم‌ها محاسبه شد، با مرتب‌سازی عناصر برحسب فاصله اقلیدسی، از میان k همسایه، نمونه‌ای که از میان k همسایه دارای کمترین مقدار است به نمونه ناشناخته داده می‌شود.

در سیستم‌های توصیه‌گر تاریخچه علاقه‌مندی‌های کاربر به آیتم‌های خریداری شده در ماتریس رتبه ذخیره می‌شود. سطرهای این ماتریس نمایانگر کاربران و ستون‌های این ماتریس نشانگر آیتم‌های خریداری شده توسط کاربران هستند. هر عنصر این ماتریس، نشان‌دهنده‌ی میزان علاقه یک کاربر خاص به آیتم خاص می‌باشد. میزان علاقه یک کاربر به یک آیتم خاص به صورت صریح و یا ضمنی می‌تواند بدست بیاید، یعنی کاربر به آیتم‌هایی که انتخاب کرده یا به صورت صریح رتبه داده است یا به صورت خودکار به کالای خریداری شده توسط سیستم رتبه اختصاص داده شده است. در این ماتریس هر کاربر تنها به تعداد محدودی آیتم انتخاب شده رتبه داده است و میزان علاقه‌مندی خود را به آن آیتم‌های مورد نظر مشخص نموده است و نظر آن نسبت به بسیاری از آیتم‌ها مشخص نیست. در شکل(۱) مراحل مدل پیشنهادی نشان داده شده است.



شکل ۱: مراحل مدل پیشنهادی برمبنای ساختار فلوچارتی

۱،۳. مرحله شباهت

در این مرحله با ارزیابی روابط ماتریس کاربر- آیتم باید زیر مجموعه‌ای از نزدیک‌ترین کاربران به کاربر جدید (شبیه‌ترین کاربران به کاربر جدید) تشخیص داده می‌شوند. در این مرحله، از فاصله اقلیدسی برمبنای امتیازات کاربر قبلی و کاربر جدید برای بدست آوردن میزان شباهت استفاده می‌شود. به منظور ارزیابی میزان شباهت مقدار امتیاز آیتم‌ها محاسبه می‌شوند اگر مقدار بدست آمده نزدیک به صفر باشد آنگاه شباهت بیشتر است. برای مثال اگر کاربر جدید به یکی از آیتم‌ها امتیاز ۴ داده باشد، این مقدار با مقدار امتیازهای قبلی محاسبه می‌شود و مقدار فاصله اقلیدسی آیتم‌هایی که با آیتم جدید برابر با صفر باشد به عنوان تشابه انتخاب می‌شوند.

اگر در محاسبه فاصله بین کاربران میزان تشابه وجود نداشته باشد آنگاه همسایه‌های دیگر بررسی می‌شوند. داده‌های ورودی در فیلترینگ مشارکتی معمولاً به عنوان یک ماتریس ارزیابی $R(m,n)$ از کاربران m×n بیان می‌شود، به‌طوریکه سطر m نشانگر کاربران M است و ستون n به معنای آیتم N است. Rij مقدار رتبه آیتم j توسط کاربر i است. تعریف رابطه بین کاربران-آیتم‌ها به صورت معادله(۴) است. هر کاربر به آیتم‌ها یک رتبه داده است که به صورت عدد ذخیره شده است.

$$R_{mn} = \begin{bmatrix} R_{11} & \cdots & R_{1n} \\ \vdots & \ddots & \vdots \\ R_{m1} & \cdots & R_{mn} \end{bmatrix} \qquad (۴)$$

در یک سیستم توصیه‌گر با n کاربر و m آیتم، مجموعه کاربران را به صورت U و مجموعه آیتم‌ها را با I نشان می‌دهیم. ماتریس کاربر-آیتم از ورودی‌های اصلی یک سیستم توصیه‌گر می‌باشد. این ماتریس شامل رتبه‌ی کاربران به آیتم‌های موجود در سیستم است. یک فاکتور مهم در الگوریتم‌های فیلترینگ مشارکتی انتخاب مناسب یک تابع شباهت می‌باشد. زیرا دقت و عملکرد توصیه را تحت تاثیر قرار می‌دهد. الگوریتم نزدیک‌ترین همسایه در فیلترینگ مشارکتی مبتنی بر حافظه براساس معیار فاصله، شباهت را محاسبه می‌کند. این معیار برمبنای مقادیری که توسط کاربران به هر آیتم داده شده است تعریف می‌شود.

۲،۳. گام پیش‌بینی

بعد از تولید ماتریس مجاورتی با استفاده از امتیازهای متعلق به هر آیتم که توسط کاربران همسایه امتیاز داده شده‌اند، امتیاز هر آیتم برای کاربر جدید محاسبه می‌شود و در قالب پیش‌بینی ارائه می‌گردد. هدف از این مرحله پیش‌بینی امتیازهای هر آیتم بدون امتیاز برای کاربران جدید با استفاده از فیلترینگ مشارکتی مبتنی بر آیتم است. امتیازهای جدید پیش‌بینی شده به عنوان امتیازی برای آیتم‌های بدون رتبه در نظر گرفته می‌شوند و به عنوان نتایج نهایی ثبت می‌شوند. تعریف پیش‌بینی طبق معادله(۵) انجام می‌شود. پس از به‌دست آوردن یک مجموعه از همسایه‌ها و تعریف فیلم‌های امتیاز داده شده توسط همسایه‌ها می‌توان لیستی از توصیه‌ها را به کاربر جدید پیشنهاد داد که توسط وی امتیازبندی نشده‌اند. همچنین رأی آن فیلم‌ها باید پیش‌بینی گردد که کاربر جدید با انتخاب آیتم‌های دلخواه به ثبت رای و پیش‌بینی کمک می‌کند. رأی برای فیلم i برای کاربر جدید u به منظور پیش‌بینی طبق معادله(۵) تعریف می‌شود.

$$\boldsymbol{predection_{u,i}} = \bar{v}_u + k\sum_{j=1}^{n} d(u,j)(v_{j,i} - \bar{v}_j) \qquad (۵)$$

پارامتر $\bar{v}_u$ میانگین امتیاز از کاربر u، پارامتر k فاکتور نرمالسازی برای محدوده بهینه که بین صفر و یک است. پارامتر d فاصله اقلیدسی، پارامتر n تعداد همسایه‌ها، پارامتر $v_{j,i}$ مقدار امتیاز واقعی از همسایه jام برای آیتم iام است.

۳،۳. معیارهای ارزیابی

باید به منظور کارایی و درستی مدل، تحلیل و ارزیابی سیستم‌های توصیه‌گر برمبنای معیارهای مختلف انجام شود. برای ارزیابی سیستم‌های توصیه‌گر معیارهای گوناگونی در منابع مختلف تعریف شده است. از جمله معمولی‌ترین معیارهای ارزیابی سیستم‌های توصیه‌گر معیارهای دقت، فراخوانی و معیار F1 می‌باشد. دقت، یکی از مهمترین اهداف طراحی در سیستم‌های توصیه‌گر می‌باشد. در ادامه معیارهای مورد نظر را توضیح خواهیم داد و در این تحقیق از معیارهای بیان شده برای مقایسه عملکرد و کارایی مدل پیشنهادی استفاده خواهیم کرد.

۱،۳،۳. میانگین خطای مطلق

به‌منظور اندازه‌گیری دقت نتایج سیستم‌های توصیه‌گر پیشنهادی، معمولا از رایج‌ترین معیارهای پیش‌بینی خطا استفاده می‌شود که از

جمله آن‌ها میانگین خطای مطلق یا همان MAE[2] و معیارهای وابسته به آن مثل میانگین خطای مربعات ریشه یا همان RMSE[3] می‌باشد. این دو معیار، خطای پیش‌بینی کاربر را محاسبه می‌کنند. میانگین خطای مطلق، میانگین انحراف بین امتیاز پیش‌بینی شده و امتیاز واقعی کاربر را محاسبه می‌کند. از جمله مزیت‌های MAE این است که نحوه‌ی محاسبه‌ی آن ساده و قابل فهم بوده و همچنین دارای ویژگی‌های آماری می‌باشد. هر چه مقدار MAE کمتر باشد، پیش‌بینی دقیق‌تر خواهد بود. فاکتور MAE طبق معادله(۶) تعریف می‌شود.

$$\boldsymbol{MAE = \frac{1}{N}\sum_{u,i}|p_{u,i} - r_{u,i}|} \qquad (۶)$$

که در آن $p_{u,i}$ و $r_{u,i}$ به‌ترتیب امتیاز پیش‌بینی شده و امتیاز واقعی کاربر u برای آیتم i و N تعداد کل امتیازات به‌مجموعه‌ی آیتم‌ها می‌باشد. همچنین ریشه میانگین خطای مطلق به‌صورت معادله(۷) تعریف می‌شود.

$$\boldsymbol{RMSE = \sqrt{\frac{1}{N}\sum_{u,i}(p_{u,i} - r_{u,i})^2}} \qquad (۷)$$

۲،۳،۳. معیارهای صحت

هدف از معیارهای صحت و جامعیت محاسبه‌ی میزان درست یا نادرست بودن توصیه‌های صورت گرفته می‌باشد. معیار صحت درستی پیشنهادها را اندازه‌گیری می‌کند. در واقع این معیار بیان می‌کند که از مجموعه پیشنهادهایی که به‌کاربر داده می‌شود چند درصد درست است بنابراین مقدار بالای این معیار نشان‌دهنده‌ی تعداد اشتباهات کمتر می‌باشد معیار مورد نظر به‌صورت معادله(۸) تعریف می‌شود.

$$Precision = \frac{TP}{TP+FP} \qquad (۸)$$

در معادله(۸) پارامتر TP تعداد آیتم‌های پیشنهاد شده مورد علاقه کاربران جدید می‌باشد در حالیکه FP تعداد آیتم‌های پیشنهادی است که کاربران جدید به آن علاقه ندارند. مجموع TP و FP تعداد کل پیشنهادات داده شده به کاربران جدید می‌باشد. معیار فراخوانی نشان می‌دهد که چنددرصد از آیتم‌های مورد علاقه کاربر به وی پیشنهاد شده است. این معیار به‌صورت معادله(۹) محاسبه می‌شود. که در آن FN تعداد کاربران جدیدی است که به آیتم‌ها علاقه دارند ولی به آن‌ها پیشنهاد نشده است. و مجموع TN و FN مجموعه‌ای از آیتم‌های مورد علاقه کاربران جدید است که باید به آن‌ها پیشنهاد شود.

$$\boldsymbol{Recall = \frac{TP}{TP + FN}} \qquad (۹)$$

در معادله(۹) بررسی می‌شود که آیا کاربر از پیشنهاداتی که سیستم به وی داده است، راضی می‌باشد یا خیر. به‌عبارتی دیگر درصد رضایت کاربر از آیتم‌های پیشنهاد شده به وی توسط معیار فراخوانی، ارزیابی می‌شود. دو معیار دقت و فراخوانی در تقابل با یکدیگر تغییر می‌کنند. ویژگی دو معیار دقت و فراخوانی با هم متفاوت است. غالباً افزایش تعداد کل آیتم‌ها باعث افزایش در نتایج فراخوانی و کاهش مقدار دقت می‌شود. برای یافتن توازن مناسب بین مقادیر دقت و فراخوانی معیار دیگری به نام معیار F1 که معمولا با نام معیارF اندازه‌گیری بیان می‌شود تعریف شده است. این معیار مقادیر دقت و فراخوانی را در یک معیار ترکیب می‌کند و به هر یک از آن‌ها وزن‌های مساوی تخصیص می‌دهد. در واقع این معیار، معیارهای دقت و فراخوانی را به‌وسیله‌ی میانگین هارمونیک آن‌ها با هم ترکیب می‌کند. مقدار F1 به‌شدت به‌سمت مقادیر کمتر از دقت و فراخوانی تمایل دارد. هر چه مقدار F1 به ۱۰۰ درصد نزدیک‌تر باشد حکایت از دقت بیشتر خروجی و نزدیک بودن آن به خواست کاربر دارد.

$$\boldsymbol{F1 = \frac{2 * Precision * Recall}{Precision + Recall}} \qquad (۱۰)$$

۴. ارزیابی و نتایج

برای آزمایش مدل پیشنهادی از نرم‌افزار متلب ۲۰۱۶ استفاده شده است. نرم‌افزار متلب که توسط کمپانی مس‌ورک توسعه یافته، به‌عنوان یکی از پیشرفته‌ترین نرم‌افزارهای علمی شناخته می‌شود که می‌تواند برای محاسبه عددی داده‌ها، توسعه الگوریتم‌ها و تجزیه و تحلیل داده‌ها مورد استفاده قرار گیرد. حجم مجموعه داده‌ها بسیار بالا است و نیاز به محاسبات بالایی دارد. برای راه‌اندازی برنامه نیاز به مواردی همانند خواندن فایل ورودی، تشکیل ماتریس کاربر-آیتم، محاسبه متوسط امتیازهایی که یک کاربر به آیتم‌ها داده است، و به‌دست آوردن لیست آیتم‌هایی که کاربر به آنها امتیاز داده است. در این پژوهش، آزمایش‌ها برروی سیستم عامل ویندوز ۸/۱ و با پردازنده‌ای ۶۴ بیتی و دو هسته‌ای، هر هسته ۲/۵ گیگاهرتز و حافظه ۴ گیگابایتی اجرا شده است.

۱،۴. ارزیابی و نتایج: معیارهای صحت

در این مقاله جهت ارزیابی نتایج شبیه‌سازی از مجموعه داده Movielens استفاده شده است [۲۱]. براساس بررسی‌های انجام شده، مجموعه داده MovieLens 100K شامل ۱۰۰۰۰۰ رکورد رتبه‌ای برای ۱۶۸۲ آیتم است که توسط ۹۴۳ کاربر ثبت شده است و مقیاس رتبه‌دهی به صورت اعداد ۱ تا ۵ است. نرخ پراکندگی برابر با ۹۵/۶۹ درصد و حداکثر تعداد رای‌های کاربران برابر با ۷۳۷ و میانگین تعداد رای‌های کاربران برابر با ۱۰۶ است. در ابتدا، تاثیر تعداد همسایگی بر معیارهای صحت بررسی شده است. ابتدا مقادیر دقت، فراخوانی و F1 برروی مجموعه داده Movielens با استفاده از تعداد همسایه‌ها (۵، ۱۰، ۱۵، ۲۰، ۲۵، ۳۰، ۴۰ و ۵۰) محاسبه شده است. نتایج در جدول(۱) نشان داده شده است که بهترین دقت متعلق به مدل پیشنهادی است.

دقت مدل پیشنهادی با ۱۰ همسایه برابر با ۰/۵۳۱۴ و با ۵۰ همسایه برابر با ۰/۴۰۱۱ است. فراخوانی مدل پیشنهادی با ۱۰ همسایه برابر با ۰/۵۴۱۷ و با ۵۰ همسایه برابر با ۰/۴۱۲۶ است. نتایج نشان می‌دهد اگر تعداد همسایه‌ها بیشتر شود درصد دقت و فراخوانی کمتر می‌شود، به

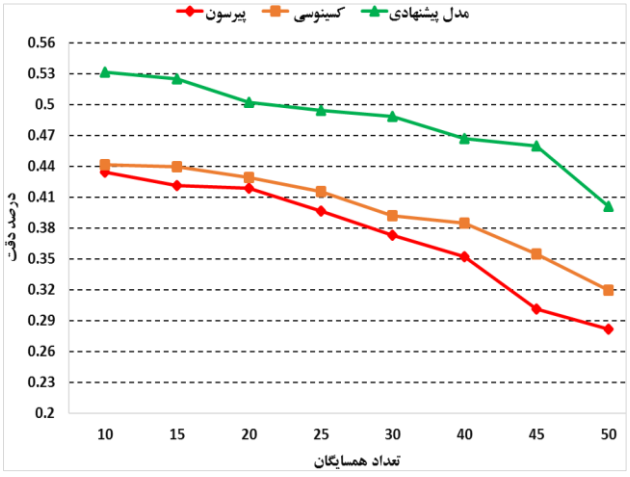
دلیل اینکه فاصله بین آیتم‌ها بیشتر می‌شود و تشخیص بهترین آیتم برای همسایگی ممکن است با اشتباه صورت بگیرد.

نتایج جدول(۱) نشان می‌دهد که دقت شباهت پیرسون با ۱۰ همسایه برابر با ۰/۴۳۴۱ و با ۵۰ همسایه برابر با ۰/۲۸۱۴ است. فراخوانی شباهت پیرسون با ۱۰ همسایه برابر با ۰/۴۳۶۴ و با ۵۰ همسایه برابر با ۰/۲۸۵۲ است. همچنین دقت شباهت کسینوسی با ۱۰ همسایه برابر با ۰/۴۴۱۳ و با ۵۰ همسایه برابر با ۰/۳۱۹۶ است. فراخوانی شباهت کسینوسی با ۱۰ همسایه برابر با ۰/۴۴۷۲ و با ۵۰ همسایه برابر با ۰/۳۲۰۸ است.

جدول ۱: ارزیابی صحت در مدل پیشنهادی برحسب تعداد همسایه‌ها

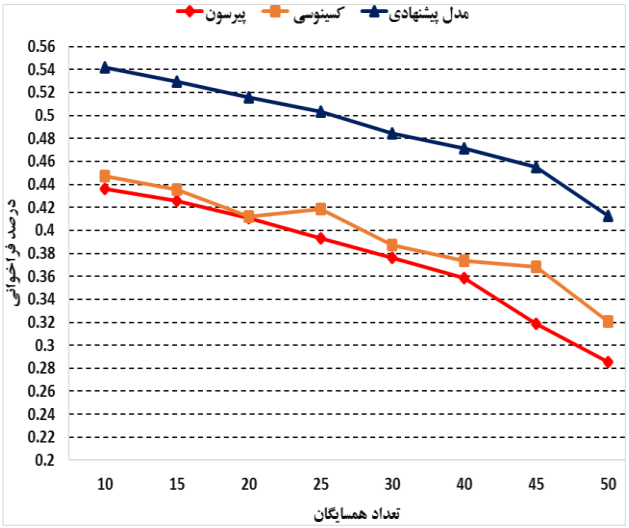
مدل‌ها	معیارها	۱۰	۱۵	۲۰	۲۵	۳۰	۴۰	۴۵	۵۰
شباهت پیرسون	دقت	۰/۴۳۴۱	۰/۴۲۱۲	۰/۴۱۸۵	۰/۳۹۶۲	۰/۳۷۳۱	۰/۳۵۲۱	۰/۳۰۱۴	۰/۲۸۱۴
	فراخوانی	۰/۴۳۶۴	۰/۴۲۵۸	۰/۴۱۱۰	۰/۳۹۳۱	۰/۳۷۶۱	۰/۳۵۸۴	۰/۳۱۸۶	۰/۲۸۵۲
	F1	۰/۴۴۰۲	۰/۴۲۶۵	۰/۴۱۵۴	۰/۳۹۵۲	۰/۳۷۲۸	۰/۳۵۴۷	۰/۳۰۹۸	۰/۲۸۴۱
شباهت کسینوسی	دقت	۰/۴۴۱۳	۰/۴۳۹۷	۰/۴۲۸۹	۰/۴۱۵۲	۰/۳۹۱۶	۰/۳۸۴۶	۰/۳۵۴۷	۰/۳۱۹۶
	فراخوانی	۰/۴۴۷۲	۰/۴۳۵۶	۰/۴۱۲۳	۰/۴۱۸۶	۰/۳۸۷۴	۰/۳۷۳۲	۰/۳۶۸۴	۰/۳۲۰۸
	F1	۰/۴۴۴۲	۰/۴۳۶۵	۰/۴۱۹۵	۰/۴۱۷۲	۰/۳۸۹۵	۰/۳۷۸۵	۰/۳۵۹۶	۰/۳۱۲۵
مدل پیشنهادی	دقت	۰/۵۳۱۴	۰/۵۲۴۹	۰/۵۰۲۱	۰/۴۹۴۵	۰/۴۸۸۴	۰/۴۶۷۱	۰/۴۵۹۴	۰/۴۰۱۱
	فراخوانی	۰/۵۴۱۷	۰/۵۲۹۴	۰/۵۱۵۶	۰/۵۰۳۲	۰/۴۸۴۷	۰/۴۷۱۷	۰/۴۵۴۸	۰/۴۱۲۶
	F1	۰/۵۳۶۵	۰/۵۲۴۹	۰/۵۰۳۵	۰/۴۹۷۵	۰/۴۸۳۵	۰/۴۶۷۳	۰/۴۵۳۵	۰/۴۰۶۸

در شکل(۲) نمودار مقایسه مدل پیشنهادی با مدل‌های پیرسون و کسینوسی برمبنای معیار دقت نشان داده شده است. نتایج در شکل(۲) نشان می‌دهد که بهترین دقت را مدل پیشنهادی و با تعداد ۱۰ همسایه ارائه داده است. معیار دقت در پیرسون و کسینوسی کمتر است و این بیانگر این است که مدل پیشنهادی مقدار خطای کمتری داشته است و همسایه را بهتر از لحاظ تشابه تشخیص داده است.



شکل۲ : نمودار مقایسه مدل پیشنهادی با مدل‌های پیرسون و کسینوسی برمبنای معیار دقت

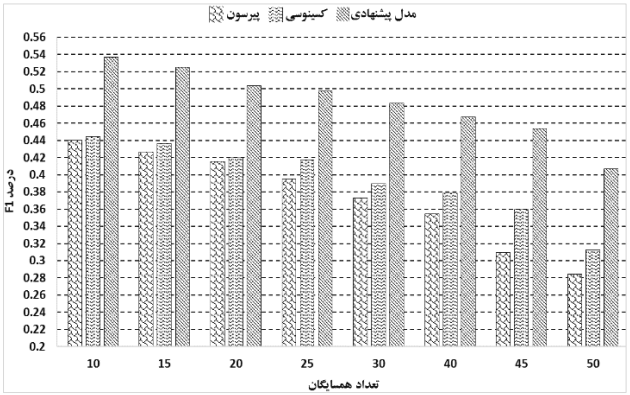
در شکل(۳) نمودار مقایسه مدل پیشنهادی با مدل‌های پیرسون و کسینوسی برمبنای معیار فراخوانی نشان داده شده است. نمودار شکل(۳) نشان می‌دهد که با افزایش تعداد همسایه‌ها در مدل پیشنهادی و مدل‌های دیگر، درصد فراخوانی کاهش پیدا می‌کند. این بدان دلیل است که اگر تعداد آیتم‌های پیش‌بینی زیاد باشد آنگاه تشخیص کاربران مشابه برای کاربر جدید به دلیل میانگین رای‌های پایین ممکن نیست و مقدار فراخوانی کمتر می‌شود. اما درصد فراخوانی در شباهت پیرسون و کسینوسی در همان اوایل در مقایسه با مدل پیشنهادی کمتر است. این دو مدل برای محاسبه تشابه از محاسبه فاصله مستقیم بین کاربران قبلی و کاربر جدید استفاده نمی‌کنند و به دلیل استفاده از فاکتورهای مختلف از آیتم مشابه دور می‌شوند و تشخیص خطا بیشتر می‌شود.



شکل ۳: نمودار مقایسه مدل پیشنهادی با مدل‌های پیرسون و کسینوسی برمبنای معیار فراخوانی

در شکل(۴) نمودار مقایسه مدل پیشنهادی با مدل‌های پیرسون و کسینوسی برمبنای معیار F1 نشان داده شده است. نتایج شکل(۴) نشان

می‌دهد که معیار F1 در مدل پیشنهادی در مقایسه با شباهت پیرسون و کسینوسی بیشتر است. معیار F1 در مدل پیشنهادی در مقایسه با شباهت پیرسون و شباهت کسینوسی بترتیب ۰/۱۲۲۷ و ۰/۰۹۴۳ درصد بهبود یافته است.



شکل ۴: نمودار مقایسه مدل پیشنهادی با مدل‌های پیرسون و کسینوسی برمبنای معیار F1

۲،۴. ارزیابی و نتایج: معیارهای خطا

جدول(۲) نشان می‌دهد که با افزایش تعداد همسایه‌ها مقدار MAEو RMSE کمتر شده است. با افزایش تعداد همسایه‌ها، مقدار MAE در مدل پیشنهادی کاهش زیادی داشته است که بیانگر افزایش قابل توجه در صحت پیش‌بینی است. نتایج نشان می‌دهد که مقدار MAEو RMSE در مدل پیشنهادی برای حالت ۵۰ همسایه برابر با ۰/۷۳۱۵ و ۰/۹۵۲۱ می‌باشد. مقدار MAEو RMSE در شباهت پیرسون برای حالت ۵۰ همسایه برابر با ۰/۹۲۴۱ و ۰/۹۷۱۶ می‌باشد. مقدار MAEو RMSE در شباهت کسینوسی برای حالت ۵۰ همسایه برابر با ۰/۹۱۴۸ و ۰/۹۶۸۱ می‌باشد. به طور کلی همانطور که در جدول(۲) نشان داده شده است مدل پیشنهادی توانسته است به مقادیر خطای کمتری در مقایسه با روش‌های دیگر دست یابد. مدل پیشنهادی مقدار خطای MAE را در مقایسه با پیرسون در حدود ۰/۱۹۲۶ کمتر کرده است.

جدول ۲: ارزیابی خطا در مدل پیشنهادی برحسب تعداد همسایه‌ها

مدل‌ها	معیارها	۱۰	۱۵	۲۰	۲۵	۳۰	۴۰	۴۵	۵۰
شباهت پیرسون	MAE	۰/۹۵۷۴	۰/۹۵۲۸	۰/۹۴۶۸	۰/۹۴۰۷	۰/۹۳۴۸	۰/۹۳۱۵	۰/۹۲۱۷	۰/۹۲۴۱
	RMSE	۱/۰۵	۱/۰۳	۰/۹۹۹۳	۰/۹۹۷۱	۰/۹۸۷۴	۰/۹۷۵۸	۰/۹۷۳۴	۰/۹۷۱۶
شباهت کسینوسی	MAE	۰/۹۵۱۹	۰/۹۴۷۶	۰/۹۴۱۱	۰/۹۳۹۴	۰/۹۳۱۴	۰/۹۲۸۸	۰/۹۲۶۱	۰/۹۱۴۸
	RMSE	۰/۹۹۷۹	۰/۹۹۵۸	۰/۹۸۸۶	۰/۹۸۶۲	۰/۹۸۱۴	۰/۹۷۷۸	۰/۹۷۰۵	۰/۹۶۸۱
مدل پیشنهادی	MAE	۰/۸۸۴۱	۰/۸۶۲۱	۰/۸۵۱۴	۰/۸۳۷۶	۰/۸۱۴۹	۰/۷۹۳۵	۰/۷۶۱۴	۰/۷۳۱۵
	RMSE	۰/۹۸۴۳	۰/۹۸۲۲	۰/۹۷۸۵	۰/۹۷۳۹	۰/۹۶۹۱	۰/۹۶۴۸	۰/۹۵۷۹	۰/۹۵۲۱

۵. نتیجه‌گیری

در حوزه سیستم‌های توصیه‌گر، تعدادی کاربر وجود دارد که نظرات خود را نسبت به تعدادی آیتم اعلام می‌نمایند. با فرض اینکه یک سیستم توصیه‌گر دارای m کاربر و n آیتم باشد و رای‌های کاربران به آیتم‌ها در محدوده $r_{min}, \ldots, r_{max}$ باشد، که در آن r_{min} و r_{max} به ترتیب بیانگر کوچک‌ترین و بزرگ‌ترین مقدار رای‌ها می‌باشند. از نماد U برای نمایش مجموعه کاربران و I برای نمایش مجموعه آیتم‌ها استفاده می‌شود. نظر هر کاربر نسبت به یک آیتم رای نامیده می‌شود که معمولاً به صورت یک سه‌گانه $(u, i, r_{u,i})$ نشان داده می‌شود که در آن u معرف یک کاربر، i معرف یک آیتم و $r_{u,i}$ نشان‌دهنده ارزش رای داده شده توسط کاربر u به آیتم i می‌باشد.

مسئله‌ی مبتنی بر ساختار امتیازدهی کاربران به مجموعه‌ای از آیتم‌ها را سیستم‌های توصیه‌گر می‌گویند. امتیاز داده شده هر کاربر به هر یک از آیتم‌ها، نشان‌دهنده نرخ علاقه کاربر به آن آیتم‌ها و تحت عنوان رای کاربر به آن آیتم‌ها محسوب می‌شود. لذا در فرم کلی، اینگونه مسائل را می‌توان مسئله پیش‌بینی امتیاز برای آیتم‌هایی در نظر گرفت که هنوز توسط کاربران دیده نشده‌اند. فیلترینگ مشارکتی، به طور معمول یک ماتریس رای که شامل نظرات کاربران نسبت به آیتم‌ها است و پیش‌بینی پیشنهادها بر اساس علائق کاربران مشابه تولید و ارائه می‌شوند. در مدل پیشنهادی، برای پیدا کردن شباهت میان دو کاربر، تابعی جدید برمبنای فاصله اقلیدسی پیشنهاد شد. تشخیص این شباهت، بر اساس نزدیکی امتیازهای دو کاربر عمل می‌کند. نزدیکی بین دو رای می‌تواند تاثیر زیادی در نتیجه نهایی داشته باشد و موجب کاهش خطا می‌شود. مدل پیشنهادی برروی مجموعه داده Movielens ارزیابی شد و توسط معیارهای صحت و خطا مورد سنجش و تست قرار گرفت. درصد معیارهای صحت نشان دادند که مدل پیشنهادی دارای درصد دقت بهتری در مقایسه با پیرسون و کسینوسی است و همچنین مقدار خطا در مدل پیشنهادی کمتر بود.

برای کارهای آینده، می‌توان از خوشه‌بندی فازی به منظور تشخیص کاربران مشابه استفاده کرد. یکی از روش‌هایی که برای کنار هم قرار دادن کاربران مشابه استفاده می‌شود، خوشه‌بندی است. با استفاده از خوشه‌بندی می‌توان کاربران مشابه به هم را در یک خوشه و کاربرانی که رفتار شبیه به هم ندارند را در خوشه‌های متفاوت قرار داد. در خوشه‌بندی فازی، هر کاربر می‌تواند مربوط به چندین خوشه باشد ولی درجه یا میزان ارجحیت آن کاربر به هر خوشه، بستگی به میزان شباهت آن کاربر با مرکز خوشه مورد نظر دارد. در این حالت تعداد زیادی از کاربران مشابه با درجه‌ای مشخص، به کاربر جدید تشابه دارند.

۶. مراجع

[1] M. Duma, B. Twala, Optimising latent features using artificial immune system in collaborative filtering for recommender systems, Applied Soft Computing, Volume 71, pp. 183-198, 2018.

[2] S. Jiang, S.C. Fang, Q. An, J.E. Lavery, A sub-one quasi-norm-based similarity measure for collaborative filtering in recommender systems, Information Sciences, Vol. 487, pp. 142-155, 2019.

[3] D. Valcarce, A. Landin, J. Parapar, A. Barreiro, Collaborative filtering embeddings for memory-based recommender systems, Engineering Applications of Artificial Intelligence, Vol. 85, pp. 347-356, 2019.

[4] F. Zhang, T. Gong, V.E. Lee, G. Zhao, Guangzhi Qu, Fast algorithms to evaluate collaborative filtering recommender systems, Knowledge-Based Systems, Vol. 96, pp. 96-103, 2016.

[5] T. Mohammadpour, A.M. Bidgoli, R. Enayatifar, H.H.S. Javadi, Efficient clustering in collaborative filtering recommender system: Hybrid method based on genetic algorithm and gravitational emulation local search algorithm, Genomics, In press, corrected proof, Available online 3 January 2019

[6] H. Parvin, P. Moradi, S. Esmaeili, TCFACO: Trust-aware collaborative filtering method based on ant colony optimization, Expert Systems with Applications, Vol. 118, pp. 152-168, 2019.

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

[8] M. Wasid, V. Kant, A Particle Swarm Approach to Collaborative Filtering based Recommender Systems through Fuzzy Features, Procedia Computer Science, Vol. 54, pp. 440-448, 2015.

[9] C. Ju and C. Xu, A New Collaborative Recommendation Approach Based on Users Clustering Using Artificial Bee Colony Algorithm, Hindawi Publishing Corporation, The Scientific World Journal, Article ID 869658, pp. 1-9, 2013.

[10] Y. Ar, E. Bostanci, A genetic algorithm solution to the collaborative filtering problem, Expert Systems with Applications, Vol. 61, pp. 122-128, 2016.

[11] H. Koohi, K. Kiani, User based Collaborative Filtering using fuzzy C-means, Measurement, Vol. 91, pp. 134-139, 2016.

[12] S.H.S. Chee, J. Han, K. Wang, RecTree: An Efficient Collaborative Filtering Method, International Conference on Data Warehousing and Knowledge Discovery, DaWaK 2001: Data Warehousing and Knowledge Discovery, pp. 141-151, 2001.

[13] S. Yadav, Vikesh, Shreyam, S. Nagpal, An Improved Collaborative Filtering Based Recommender System using Bat Algorithm, Procedia Computer Science, Vol. 132, 1795-1803, 2018.

[14] J.N. Ruskin, V. Fuster, Abstracts of original contributions: 42nd Annual Scientific Session, Journal of the American College of Cardiology, Vol. 21, Issue 2, pp. 17a-488a, 1993

[15] Z. Wang, X. Yu, N. Feng, Z. Wang, an improved collaborative movie recomme ndation system using computational intelligence, Journal of Visual Languages and Computing, Vol. 25, pp. 667–675, 2014.

[16] G.M. Dakhel, M. Mahdavi, Providing an Effective Collaborative Filtering Algorithm Based on Distance Measures and Neighbors' Voting, International Journal of Computer Information Systems and Industrial Management Applications, Vol. 5, pp. 524-531, 2013.

[17] R. Katarya, O.P. Verma, Effectual recommendations using artificial algae algorithm and fuzzy c-mean, Swarm and Evolutionary Computation, Vol. 36, pp. 52-61, 2017.

[18] B. Alhijawi, Y. Kilani, A collaborative fitering recommender system using genetic algorithm, Information Processing and Management, Vil. 57, pp. 1-21, 2020

[19] LC. Cheng, HA. Wang, A fuzzy recommender system based on the integration of subjective preferences and objective information, Applied Soft Computing, Vol. 18, pp. 290-301, 2014.

[20] Y. Ho, S. Fong, Z. Yan, On Improving GA-based Collaborative Filtering for Online Recommender, pp. 1-7, 2007. https://grouplens.org/datasets/movielens/100k/

زیرنویس

1. Genetic-based Recommender System (BLIGA)
2. Mean Absolute Error (MAE)
3. Root Mean Square Error (RMSE)