علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۵، شماره ۱، ۱۳۹۶
صفحات ۴۰- ۵۴
مقاله عادی

نشریه علوم رایانش و فناوری اطلاعات

یک معیار شباهت نوین پالایش مشارکتی در سامانه‌های توصیه‌گر

مجتبی کاظمی ساسان حسینعلی‌زاده

دانشکده برق ـ رایانه و فناوری اطلاعات، دانشگاه آزاد اسلامی واحد قزوین، قزوین، ایران

چکیده

سامانه‌های توصیه‌گر به سه دسته پالایش جمعیت شناختی، پالایش مبتنی بر محتوا و پالایش مشارکتی تقسیم‌بندی می‌گردند. پالایش مشارکتی مبتنی بر همسایگان به‌عنوان یکی از مهم‌ترین کلاس‌های پالایش مشارکتی، کاربرد گسترده‌ای در حوزه تجاری را داراست. کلید این رویکرد در یافتن کاربران و یا کالاهایی مشابه براساس ماتریس امتیازات کاربر- کالا بوده تا بتواند توصیه‌های مناسبی برای کاربران فراهم نماید. در این مقاله به‌منظور محاسبه شباهت میان کاربران، معیار مشابهت جدیدی را براساس همسایگان کاربران ارائه داده‌ایم تا عملکرد توصیه‌ها را زمانی که تعداد امتیازات کمی در دسترس باشد، بهبود بخشد. از این‌رو از رویکرد احتمالاتی برای مدل‌سازی معیار مشابهت پیشنهادی میان دو کاربر پیشنهاد شده است. به‌منظور نشان دادن اثربخشی معیار، عملکرد معیارهای مشابهت سنتی و مدرن را با معیار مشابهت پیشنهادی مقایسه کرده‌ایم. نتایج توصیه‌های صورت گرفته براساس معیارهای ارزیابی مختلف نشان‌دهنده این است که معیار مشابهت پیشنهادی عملکرد بهتری نسبت به دیگر معیارهای مشابهت در داده‌های پراکنده داراست.

کلمات کلیدی: سامانه‌های توصیه‌گر، پالایش مشارکتی، پالایش مشارکتی مبتنی بر همسایگان، معیار مشابهت کاربران، داده‌های پراکنده.

۱- مقدمه

امروزه استفاده گسترده از اینترنت، زیرساخت لازم را برای توسعه سریع تجارت الکترونیک کرده است [۱]. در این میان، ارائه سرویس‌های شخصی به مشتریان از معروف‌ترین و پرکاربردترین خدماتی است که در وب‌سایت‌های فعال در تجارت الکترونیک جهت ارتباط بهتر، جذب و حفظ مشتریان بکار گرفته می‌شود. سامانه‌های توصیه‌گر[۱] نقش مهمی را در این راستا جهت ارائه پیشنهادها به کاربران براساس نیازها و علایق آن‌ها ایفا می‌کنند [۲]. در واقع این سامانه‌ها به کاربران در انتخاب کالای موردنظرشان از میان فهرست طولانی از کالاهای موجود در یک فروشگاه مجازی کمک می‌کنند.

سامانه‌های توصیه‌گر با جمع‌آوری اطلاعات کاربران، علایق و ترجیحات آنها نسبت به کالاهای مختلف (از قبیل فیلم‌ها، آهنگ‌ها، کتاب‌ها، جک‌ها، گجت‌ها، برنامه‌های کاربردی، وب‌سایت‌ها و غیره) شناسایی می‌نمایند. این اطلاعات به صورت ضمنی (به صورت نظارت بر رفتار کاربران همانند آهنگ‌های شنیده شده، برنامه‌های دانلود شده، وب‌سایت‌های بازدید شده و کتاب‌های خوانده شده و غیره) و یا به صورت صریح (با استفاده از امتیازات کاربران، یا ویژگی‌های جمعیت شناختی کاربران مانند سن، جنسیت، ملیت، میزان تحصیلات و غیره) جمع‌آوری می‌شوند [۳، ۴]. همچنین اطلاعات اجتماعی[۲] از قبیل پیروان[۳]، توییت‌ها[۴] و پست‌های وب‌سایت، اطلاعاتی هستند که عموماً در وب ۲،۰ مورد استفاده قرار می‌گیرند [۵، ۶].

براساس طیف اطلاعات مورد استفاده جهت ارائه پیشنهادها، سامانه‌های توصیه‌گر به گروه‌های مبتنی بر پالایش محتوی[۵]، پالایش جمعیت شناختی[۶]، پالایش مشارکتی[۷] و پالایش ترکیبی[۸] تقسیم‌بندی می‌شوند [۷، ۸]. در سامانه‌های توصیه‌گر مبتنی بر پالایش محتوی، کالاهای جدید براساس میزان شباهت خصوصیات و ویژگی‌هایشان با کالاهایی که قبلا کاربر انتخاب نموده، به او پیشنهاد می‌شوند [۹]. در سامانه‌های توصیه‌گر مبتنی بر پالایش جمعیت شناختی، شباهت میان کاربران براساس اطلاعاتی از قبل سن، جنسیت، شغل، میزان تحصیلات و غیره محاسبه می‌شود [۱۰]. ایده اصلی سامانه‌های پالایش جمعیت شناختی این است که گروه‌های مختلف جامعه از علایق و ترجیحات یکسانی برخوردار هستند. جمع‌آوری و سازماندهی اطلاعات کالاها و کاربران هزینه عملیاتی این دو سامانه را

افزایش می‌دهد. علاوه بر این، عدم‌تمایل بسیاری از کاربران به ارائه اطلاعات شخصی و حفظ حریم خصوصی کاربران از مهم‌ترین دغدغه‌های سامانه‌های پالایش مشارکتی محسوب می‌شود.

اما در سامانه‌های توصیه‌گر مبتنی بر پالایش مشارکتی از امتیازاتی که کاربران به کالاهای مختلف داده‌اند برای این منظور استفاده می‌گردد [۱۱]. سامانه‌های پالایش مشارکتی از رویکردهای کاربر محور، کالا محور و ترکیبی استفاده می‌کنند. در رویکردهای کاربر محور، شباهت میان هر دو کاربر براساس امتیازاتی که به کالاها داده‌اند محاسبه می‌شود، سپس پیش‌بینی امتیاز یک کالا براساس امتیازاتی که کاربران مشابه به آن کالا داده‌اند صورت می‌گیرد. اما در رویکرد کالا محور، شباهت کالاها با یکدیگر براساس امتیازاتی که کاربران به آنها داده‌اند محاسبه می‌شود، سپس برای پیش‌بینی امتیاز یک کالا برای یک کاربر از امتیازاتی که او به سایر کالاهای مشابه داده است، استفاده می‌گردد. رویکردهای ترکیبی در واقع تلفیقی از دو رویکرد مذکور می‌باشند. سامانه‌های توصیه‌گر پالایش ترکیبی نیز سعی می‌کنند با استفاده توام از رویکردهای سامانه‌های مذکور، از مزایای نسبی آن‌ها- برای رفع نقاط ضعف موجود در هر یک از آن‌ها- بهره‌مند شوند [۱۲، ۱۳].

گستردگی و تنوع کالاهایی که در یک فروشگاه الکترونیک عرضه می‌شود از یک طرف و عدم‌تمایل کاربران به اظهارنظر در مورد امتیاز کالاها از طرف دیگر چالش‌های فراوانی را برای سامانه‌های توصیه‌گر در خصوص کشف علایق و ترجیحات کاربران ایجاد نموده است [۱۴]. مشکلات تنکی[۹] و شروع سرد[۱۰] در سامانه‌های توصیه‌گر مبتنی بر پالایش مشارکتی از زمره این چالش‌ها محسوب می‌گردند. واژه تنکی در اصل از ادبیات جبرخطی برگرفته شده است و به ماتریسی اطلاق می‌شود که عمده درایه‌های آن صفر باشند [۱۴]. دلیل استفاده از آن این است که معمولا امتیازاتی که کاربران به آیتم‌های مختلف می‌دهند در قالب یک ماتریس سازماندهی می‌شود. اگر کاربر iام به کالای jام امتیازی داده باشد، این امتیاز در درایه سطر i و ستون j ماتریس درج می‌گردد، در غیر این صورت مقدار درایه برابر با صفر درنظر گرفته می‌شود. اگر تعداد امتیازات کاربران کم باشد نسبت تعداد درایه‌های غیرصفر ماتریس به درایه‌های صفر کوچک شده، ماتریس امتیازات تنک می‌گردد. مشکل شروع سرد برای کاربران یا کالاهایی رخ می‌دهد که به تازگی وارد سامانه شده‌اند و هیچگونه (یا تعداد کمی) امتیازی برای آنها ثبت نشده است.

معیارهای شباهت متنوعی- از قبیل ضریب همبستگی پیرسون، معیار کسینوسی، معیار کسینوسی تنظیم شده، ضریب جاکارد و غیره تاکنون توسط محقیق در سامانه‌های توصیه‌گر مشارکتی بکار برده شده‌اند. شباهت میان دو کاربر عمدتا براساس امتیازاتی که آنها به مجموعه‌ای از کالاهای مشترک[۱۱] داده‌اند، محاسبه می‌شود. از یک طرف، مشکلات تنکی و شروع سرد تعداد کالاهایی که توسط هر دو کاربر امتیاز داده شده‌اند را کاهش می‌دهد که این امر منجر به کاهش قابلیت اطمینان میزان مشابهت محاسبه شده برای آن دو کاربر می‌شود [۱۵]؛ اگر دو کاربر صرفا به یک کالا امتیاز داده و امتیازاتشان یکسان باشند انگاه میزان شباهت آن‌ها با دو کاربر دیگر که به ۲۰ کالا امتیاز یکسان داده‌اند برابر خواهد بود، اما درجه اطمینان حالت دوم از حالت اول بسیار بیشتر است. از طرف دیگر، این مشکلات احتمال یافتن کاربران و کالاهای مشابه را به صورت چشم‌گیری کاهش می‌دهد و با کاهش تعداد کاربران/کالاهای مشابه در روند پیش‌بینی امتیاز یک کالا- از دقت و قابلیت اطمینان پیش‌بینی به میزان قابل توجهی کاسته می‌شود [۱۶].

انگیزه‌ها و نوآوری‌های ما برای ارائه یک معیار مشابهت به شرح زیر می‌باشد:

- بجای استفاده از مجموعه کالاهای مشترک، تمام کالاها (حتی کالاهایی که کاربران به آنها امتیازی نداده‌اند) برای محاسبه معیار شباهت پیشنهادی مورد استفاده قرار می‌گیرند.
- میزان اهمیت و تاثیر کالاها در محاسبه معیار شباهت پیشنهادی با یکدیگر برابر نیست.
- ضریب اهمیت هر کالا هنگام محاسبه شباهت دو کاربر با توجه به جایگاه نسبی امتیازات آنها نسبت به امتیازات سایر کاربران تعیین می‌گردد.
- از احتمال اتفاق/اختلاف نظر دوکاربر در مورد یک کالا برای محاسبه شباهت میان آن‌ها استفاده می‌شود.

مابقی مقاله به شرح زیر ساماندهی شده است. در بخش دوم مروری بر تحقیقات انجام شده در سامانه‌های توصیه‌گر مبتنی بر پالایش مشارکتی داشته و در بخش سوم ایده اصلی معیار مشابهت پیشنهادی را بیان می‌نماییم. بخش چهارم نیز نشان خواهیم داد که معیار پیشنهادی چگونه می‌تواند به کارایی بهتر سامانه‌های توصیه‌گر منجر شود. همچنین در بخش پنجم نتیجه‌گیری و کارهای آتی مربوط به ایده پیشنهادی خود را بیان می‌نماییم.

۲- پیش‌زمینه و کارهای مرتبط

در این بخش ابتدا قاعده کلی رویکرد مبتنی بر همسایگان را با جزئیات بیان نموده و سپس به مروری بر تحقیقات انجام شده در زمینه معیارهای مشابهت پالایش مشارکتی مبتنی بر همسایگان پرداخته و معیارهای مشابهت مختلفی را به‌منظور بررسی عملکرد سامانه‌های توصیه‌گر معرفی نموده‌ایم.

۲-۱- رویکرد مبتنی بر همسایگان

رویکرد مبتنی بر همسایگان و یا رویکرد مبتنی بر حافظه به‌طور مستقیم در مجموعه مقالات توصیه‌گر گروپلنز منتشر گردیده [۱۷] و با توجه به کاربرد گسترده آن در مسائل تجاری از محبوبیت فراوانی برخوردار است [۱۸، ۱۹]. این روش با استفاده از کل پایگاه داده امتیازات، دو هدف را پیگیری می‌نماید: ۱- پیش‌بینی امتیاز برای یک کالا (محصول)، ۲- لیستی از کالاهای توصیه شده را برای کاربر هدف تهیه می‌نماید. درنظر بگیرید که $R=[r_{ui}]_{M*N}$ ماتریس امتیازات (مجموعه داده) در یک سامانه توصیه‌گر مبتنی بر پالایش مشارکتی است که در آن، هر ورودی r_{ui} بیانگر مقدار امتیاز داده شده توسط کاربر U_u به کالا I_i می‌باشد. به‌طور کلی مقدار امتیاز داده شده در یک محدوده تعیین شده (بطور مثال در مجموعه داده موویلنز بین ۱ تا ۵ است) می‌باشد. عملیات پیش‌بینی در الگوریتم‌های پالایش مشارکتی مبتنی بر همسایگان به این نحو می‌باشد که پیش‌بینی امتیاز برای i امین کالا با استفاده از اطلاعات همسایه u امین کاربر (روش کاربر محور) و یا با استفاده از اطلاعات همسایه i امین کالا (روش کالا محور) صورت می‌گیرد.

۲-۱-۱- روش‌های کاربر محور

در روش‌های کاربر محور پیش‌بینی امتیاز i امین کالا براساس u امین کاربر محاسبه می‌گردد [۱۰، ۲۰]. از همین رو در روش‌های کاربر محور ابتدا مشابهت بین کاربران هدف (که منظور همان U_u می‌باشد) و U_p، $(p=1, \ldots, M, p \neq u)$ محاسبه گردیده، سپس نزدیک‌ترین k کاربر را برای تشکیل همسایگان کاربر هدف انتخاب می‌گردد. در نهایت با استفاده از معادله (۱) یک امتیاز $(\hat{r}_{ui})$ را برای کالا iام پیش‌بینی می‌نماید.

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{k=1}^{k} s(U_u,U_k)*(r_{ki}-\bar{r}_{ku})}{\sum_{k=1}^{k}|s(U_u,U_k)|} \qquad (۱)$$

در معادله (۱)، $\bar{r}_u$ متوسط امتیاز داده شده توسط کاربر U_u است. همچنین $s(U_u, U_k)$ بیانگر مقدار مشابهت مابین کاربر هدف و همسایه kام می‌باشد. از این‌رو r_{ki} امتیاز داده شده توسط همسایه kام به کالای iام بوده و $\bar{r}_{k_u}$ متوسط امتیازات داده شده توسط kامین همسایه کاربر U_u می‌باشد.

۲-۱-۲- روش‌های کالا محور

در روش‌های کالا محور که توسط بزرگ‌ترین خرده‌فروشان (همانند وب‌سایت آمازون [۲۱]) مورداستفاده قرار می‌گیرد، ابتدا مشابهت بین کالا هدف I_i و دیگر کالاها I_j، $(j=1, \ldots, N, i \neq j)$ محاسبه گردیده، سپس k کالای که دارای بیشترین مشابهت می‌باشند، انتخاب می‌گردد. در نهایت این روش با استفاده از این k کالا و براساس معادله (۲) یک امتیاز $(\hat{r}_{ui})$ را برای کاربر هدف U_u پیش‌بینی می‌نماید [۲۲].

$$\hat{r}_{ui} = \bar{r}_i + \frac{\sum_1^k s(I_i,I_k)*(r_{uk}-\bar{r}_k)}{\sum_1^k |s(I_i,I_k)|} \quad (۲)$$

در معادله (۲)، $\bar{r}_i$ متوسط امتیاز داده شده توسط همه کاربران به کالا I_j می‌باشد. همچنین $s(I_i, I_k)$ بیانگر مقدار مشابهت بین کالا I_i و kامین کالا مشابه می‌باشد و r_{uk} امتیاز داده شده توسط کاربر فعال به k امین کالا مشابه I_i می‌باشد.

محاسبه مشابهت یکی از گام‌های حیاتی در پالایش مشارکتی مبتنی بر همسایگان بوده و بسیاری از معیارهای مشابهت در حوزه‌های مختلفی از جمله یادگیری ماشین، بازیابی اطلاعات و آمار معرفی شده است. محققان و پژوهشگران در حوزه سامانه‌های توصیه‌گر یا به‌طور مستقیم از آن‌ها بهره گرفته‌اند و یا معیارهای مشابهت جدیدی را ابداع نموده‌اند که به‌طور خلاصه ما در ادامه آن‌ها را ذکر می‌نماییم.

۲-۲- معیارهای مشابهت در پالایش مشارکتی

بیشترین بهبودهای صورت گرفته در مدل‌های پالایش مشارکتی منجر به ایجاد مدل‌های پیچیده‌تر و یا افزودن بهبودهای جدید به مدل‌های شناخته شده می‌باشد [۲۳]. به‌طور کلی پالایش مشارکتی مبتنی بر همسایگان از معیار مشابهت برای پیدا کردن همسایه‌های یک کاربر فعال و یا پیدا کردن کالاهای مشابه به کالا هدف استفاده می‌نماید. معیارهای مشابهت سنتی از قبیل ضریب همبستگی پیرسون[۱۲]، مشابهت کسینوسی و انواع مختلف آن‌ها غالباً برای محاسبه مشابهت بین یک جفت از کاربران و یا بین یک جفت از کالاها استفاده می‌شود. از دیگر محبوب‌ترین معیارهای استفاده شده در پالایش مشارکتی مبتنی بر همسایگان می‌توان به معیار کسینوس تنظیم شده[۱۳]، معیار PC مقید[۱۴] و JMSD و معیارهای مشابهت نوینی از قبیل NHSM، PIP و BCF مطابق با جدول ۱ اشاره نمود.

معیار کسینوس از محبوب‌ترین معیارها در حوزه داده‌کاوی محسوب می‌شود [۲۳]. در این معیار به‌منظور محاسبه شباهت میان دو کاربر U و V، آن‌ها را به‌عنوان بردار N بعدی امتیازات در نظر می‌گیرند، بدین ترتیب $U,V \in N_0^N$ که N_0 مجموعه‌ای از اعداد طبیعی است که صفر را نیز شامل می‌گردد. سپس مطابق جدول ۱ مقدار شباهت بین دو کاربر برابر با کسینوس زاویه مابین کاربر U و V می‌باشد. براساس معادله معیار شباهت کسینوسی r_{UI} امتیاز کاربر U به کالا I، r_{VI} بیانگر امتیاز کاربر V به کالا I و I' بیانگر مجموعه کالاهایی است که دارای امتیاز مشترک مابین کاربر U و V می‌باشد. معیار مشابهت کسینوس در پالایش مشارکتی کالا محور از محبوب‌ترین معیارها بشمار می‌آید. با این حال معیار مشابهت کسینوس در هنگام محاسبه مشابهت مابین یک جفت کالا، تفاوت مقیاس (محدوده) امتیازاتی را که توسط کاربر مهیا شده است را در نظر نمی‌گیرد. معیار مشابهت کسینوس تنظیم شده [۲۲] با توجه به مشکلاتی که در معیار مشابهت کسینوس وجود داشت مبادرت به کم کردن میانگین کاربر متناظر از امتیاز داده شده به کالا می‌نماید. از همین رو این معیار مطابق معادله درج شده در جدول ۱، به محاسبه همبستگی خطی مابین امتیازات دو کالا می‌پردازد.

ضریب همبستگی پیرسون [۲۰] در پالایش مشارکتی کاربر محور یکی دیگر از محبوب‌ترین معیارها بشمار می‌آید. معیار مشابهت پیرسون دو کاربر یا کالا را از منظری بررسی می‌نماید که آن‌ها چقدر از لحاظ خطی به یکدیگر مرتبط می‌باشند. از این‌رو ضریب همبستگی پیرسون، همبستگی کالاهایی که به‌طور مشترک توسط کاربران U و V امتیاز داده‌شده باشند را مطابق با معادله جدول ۱ محاسبه می‌نماید. مقدار عددی ضریب همبستگی پیرسون در محدوده ۱- تا ۱+ می‌باشد. مقدار عددی ۱+ بالاترین همبستگی و ۱- کمترین همبستگی مابین کالاهای مشترک توسط کاربران U و V می‌باشد. به‌طور مشابه نیز مشابهت مابین دو کالا I و J نیز با استفاده از ضریب همبستگی پیرسون به همین طریق قابل محاسبه می‌باشد. ضریب همبستگی پیرسون مقید [۲۴] نیز یکی از انواع ضریب پیرسون بوده که در آن یک مرجع مطلق (میانه در مقیاس امتیازات) بجای میانگین امتیازات کاربر متناظر مورد استفاده قرار می‌گیرد که براساس جدول ۱ محاسبه می‌گردد.

معیار مشابهت PIP [۲۵] از نوین‌ترین و مشهورترین معیارها بعد از معیارهای مشابهت رایج به‌حساب می‌آید که از طریق جدول ۱ قابل محاسبه خواهد بود. معیار PIP، سه عامل مهم به نام‌های مجاورت، تأثیر و محبوبیت را مابین امتیازات در محاسبه معیار مشابهت دو کالا (کاربر) در نظر می‌گیرد . عامل مجاورت درواقع تفاوت حسابی ساده مابین دو امتیاز داده شده به یک کالا بوده و در صورتیکه آن دو امتیاز با یکدیگر مخالف باشند، با اعمال مجازات همراه خواهد بود. مخالفت یا عدم‌مخالفت امتیازات با توجه به یک مرجع مطلق به‌طور مثال میانه محدوده امتیازات، صورت می‌گیرد. عامل تأثیر در معیار PIP، این امر را نشان می‌دهد که چقدر یک کالا ترجیح داده شده یا توسط کاربر نادیده گرفته شده است. در صورتیکه امتیازات داده شده در سمت میانه امتیازات نباشد، این ویژگی به اعمال مجازات می‌پردازد. عامل محبوبیت نیز اهمیت امتیازات داده شده را بیان می‌کند که چه مقدار از میانگین امتیازات داده شده به کالا فاصله دارد. این عامل اطلاعات سرتاسری مربوط به کالا موردعلاقه را جمع‌آوری می‌نماید. معیار PIP، این سه عامل را براساس کالاهای دارای امتیاز مشترک محاسبه کرده و در نتیجه عملکرد بهتری در ارائه توصیه‌های بهتر به کاربران را داراست.

کیم و همکارانش در [۲۶] معیار مشابهت جدیدی را برای حل مشکل شروع سرد ارائه نموده‌اند. آن‌ها یک مدل را ایجاد کرده‌اند که ابتدا به پیش‌بینی امتیاز پرداخته و سپس خطای به وجود آمده را بر روی امتیازات شناخته شده برای هر کاربر را محاسبه می‌نماید. براساس خطای محاسبه شده مدل نهایی شکل می‌گیرد. اما در هر حال این معیار نیز به‌منظور ایجاد مدل اولیه از معیارهای مشابهت رایج (مانند کسینوس و پیرسون) استفاده می‌نماید.

بباديلا و همکارانش چندین معیار مشابهت را به‌منظور حل مشکلات معیارهای مشابهت رایج ارائه کرده‌اند از این‌رو آنها در [۲۷] معیاری را ارائه داده‌اند که از ترکیب معیارهای جاکارد و مربع اختلاف میانگین به‌منظور تکمیل همدیگر به نام JMSD را معرفی کرده‌اند که از طریق جدول ۱ قابل محاسبه خواهد بود. همچنین آن‌ها در [۲۸] معیاری را براساس تکینگی ارائه نموده‌اند، که از ترکیب درصدی از امتیازات مرتبط و نامرتبط با معیار MSD کالاهای دارای امتیاز مشترک به دست می‌آید. همچنین آن‌ها معیاری در [۲۹] را ارائه نموده‌اند که از مقدار (اطلاعات) عددی امتیازات به‌خوبی توزیع امتیازات (ناپایداری) توسط جفت کاربران در هنگام محاسبه شباهت مابین آن‌ها استفاده می‌نماید. به‌منظور محاسبه اطلاعات عددی، نویسنده تعداد کالاهای دارای امتیاز مشترک با امتیاز دقیق آن، تعداد کالاهای

دارای امتیاز مشترک با محدوده امتیازی مختلف (همانند ۱، ۳، ۵) و مقدار MSD امتیازات کالاهای دارای امتیاز مشترک را محاسبه می‌نماید. معیار مشابهت جاکارد به‌منظور اختلاف امتیازات مهیا شده توسط هر دو جفت کاربران می‌پردازد.

این معیار به‌عنوان یک معیار پایه نامیده می‌شود. در نهایت این معیار پایه ترکیب گردیده تا معیاری به نام اختلاف جاکارد میانگین (MJD) شکل گیرد. همچنین از روش‌های یادگیری عصبی نیز می‌توان به‌منظور محاسبه وزن به‌عنوان یک معیار اولیه مورداستفاده قرار داد. آن‌ها نشان داده‌اند که در معیار ارزیابی میانگین خطای مطلق (MAE)، MJD مبتنی بر پالایش مشارکتی از نتایج بهتری نسبت به PIP مبتنی بر پالایش مشارکتی در تعداد همسایگان (K) بسیار زیاد برخوردار است. به هر حال تمامی این سه معیار معرفی شده از مشکل کالاهای دارای امتیاز مشترک کم رنج برده و بنابراین در کاربرد استفاده کمتری از آن‌ها به عمل می‌آید.

هایفنگ و همکارانش نوع جدیدی را از معیارهای مشابهت به نام NHSM را در [۳۰] ارائه داده‌اند تا بتواند بر مشکلات معیار PIP فائق آید. آن‌ها این موضوع را در نظر گرفتند که معیار PIP مبتنی بر پالایش مشارکتی نیاز به جریمه بیش از یک‌بار در عامل‌های تأثیر و مجاورت ندارد و تنها یک‌بار جریمه شدن کافی است. آن‌ها مطابق جدول ۱ تابع غیرخطی دیگری براساس معیار PIP را عرضه کردند که از سه عامل مجاورت، اهمیت و یکتایی به‌منظور محاسبه مشابهت استفاده می‌نماید. در نهایت این سه عامل با معیار جاکارد ترکیب می‌گردد؛ اما با این حال این عامل‌ها نیز تنها براساس کالاهای دارای امتیاز مشترک قابل محاسبه می‌باشند و امتیازات کالاهای بدون امتیاز مشترک نادیده گرفته شده‌اند.

به‌منظور بهره‌برداری از تمامی کالاهای دارای امتیاز، معیار باتاچاریا برای اولین بار در مقاله [۳۱] چاپ گردید. پاترا و همکارانش فرمول عمومی باتاچاریا را برای معیار مشابهت ارائه داده بودند که دیگر معیار مشابهت رایج می‌توانستند به‌منظور حل مشکلات شروع سرد با این فرمول ترکیب گردند.

جدول ۱- رایج‌ترین معیارهای مشابهت مورد استفاده در پالایش مشارکتی مبتنی بر همسایگان

معیار	معادله	معایب
کسینوس [۲۳]	$s(U,V)=\frac{\sum_{I\in I'}(r_{UI})(r_{VI})}{\sqrt{\sum_{I\in I'} r_{UI}^2 \sum_{I\in I'} r_{VI}^2}}$	این معیار شباهت از مشکل کالاهای دارای امتیاز مشترکمیان کاربران رنج می‌برد. همچنین علی‌رغم تفاوت‌های معنادار در امتیازات، خروجی معیار مشابهت بسیار بالا خواهد بود.
کسینوس تنظیم شده [۲۲]	$s(I,J)=\frac{\sum_{U\in U'}(r_{UI}-\bar{r}_I)(r_{UJ}-\bar{r}_J)}{\sqrt{(r_{UI}-\bar{r}_I)^2}\sqrt{(r_{UJ}-\bar{r}_J)^2}}$	در صورت اینکه مجموعه U' بسیار کم باشد، قادر به محاسبه معیار شباهت نمی باشد. از این‌رو علی‌رغم تفاوت (شباهت)های معنادار در امتیازات، خروجی معیار مشابهت بسیار بالا (کم) خواهد بود.
ضریب همبستگی پیرسون [۲۰]	$s(U,V)=\frac{\sum_{I\in I'}(r_{UI}-\bar{r}_U)(r_{VI}-\bar{r}_V)}{\sqrt{\sum_{I\in I'}(r_{UI}-\bar{r}_U)^2}\sqrt{\sum_{I\in I'}(r_{VI}-\bar{r}_V)^2}}$	این معیار شباهت از مشکل کالاهای دارای امتیاز مشترکمیان کاربران رنج می‌برد. از این‌رو علی‌رغم تفاوت (شباهت) های معنادار در امتیازات، خروجی معیار مشابهت بسیار بالا (کم) خواهد بود.
ضریب پیرسون مقید [۲۴]	$s(U,V)=\frac{\sum_{I\in I'}(r_{UI}-r_{med})(r_{VI}-r_{med})}{\sqrt{\sum_{I\in I'}(r_{UI}-r_{med})^2}\sqrt{\sum_{I\in I'}(r_{VI}-r_{med})^2}}$	این معیار مشابهت بر مشکلات ضریب همبستگی پیرسون فائق آمده اما همچنان از مشکل کالاهای دارای امتیاز مشترکمیان کاربران رنج می‌برد.
[۲۵] PIP	$s(U,V)=\sum_{k\in I'}\begin{matrix}\text{Proximity}(r_{UI},r_{VJ})\times\text{Significance}(r_{UI},r_{VJ})\\ \times\text{Singularity}(r_{UI},r_{VJ})\end{matrix}$	علی‌رغم توجه به کالاهای دارای امتیاز مشترک و مقدار عددی آنها اما نسبت امتیازات مشترک را نادیده می‌گیرد.
[۲۷] JMSD	$s(U,V)=s^{MSD}(U,V)\times s^{Jaccard}(U,V)$ s_{MSD} و $s_{Jaccard}$ به ترتیب بیانگر معیار شباهت MSD و جاکارد می‌باشند.	این معیار مشابهت بر بخشی از مشکلات MSD و Jaccard فائق آمده است. با این حال از مشکلات اطلاعات محلی و عدم به‌کارگیری تمامی امتیازات رنج می‌برد.
[۳۰] NHSM	$s(U,V)^{NHSM}=s(U,V)^{JPIP}\times s(U,V)^{URP}$ $s(U,V)^{JPSS}=s(U,V)^{PIP}\times s(U,V)^{Jaccard'}$ $s(U,V)^{URP}=1-\frac{1}{1+\exp(-\lvert\mu_U-\mu_V\rvert.\lvert\sigma_U-\sigma_V\rvert)}$	
[۳۱] BCF	$BC(p_1,p_2)=\sum_{x\in X}\sqrt{p_1(X)p_2(X)}$ $s(U,V)=\text{Jaccard}+\sum_{i\in I_U}\sum_{j\in I_V}BC(i,j)\frac{(r_{Ui}-\bar{r}_U)\times(r_{Vj}-\bar{r}_V)}{\sigma_U\sigma_V}$	

در این رویکرد در صورتیکه مشابهت بین کالاهای مشترک بدون امتیاز، حداکثر مقدار خود را داشته باشند، امتیازات جفت کالاهای مشترک بدون امتیاز نیز در محاسبه معیار مشابهت به حساب آورده می‌شوند. در غیر این صورت امتیازات آن کالاها نادیده گرفته می‌شوند. مهم‌ترین مشکل این روش این است که از تمامی امتیازات کالاهای مشترک بدون امتیاز مورد استفاده قرار نمی‌گیرد. مشابهت بین دو کاربر در صورتیکه به تعداد کالاهای کمی و یا هیچ کالا مشترکی امتیاز نداده باشند، قابل محاسبه نیست.

همچنین پاترا و همکارانش معیار مشابهت دیگری را با نام ضریب باتاچاریا ارائه داده‌اند که از فرمول باتاچاریا به‌عنوان ضریبی در دیگر معیارها استفاده می‌نماید [۳۲]. BCF علیرغم دیگر معیارهای مشابهت از تمامی امتیازات داده شده توسط کاربران بهره می‌گیرد و بدین‌ترتیب از ترکیب معیارهای مشابهت محلی و سرتاسری به منظور محاسبه شباهت میان کاربران استفاده می‌نماید. I_U و I_V بترتیب مجموعه کالاهایی است که کاربر UV و امتیاز داده‌اند که ممکن است هیچ کالای مشترکی نباشد که هر دوی کاربران به آن امتیاز داده باشند ($I_U \cap I_V \equiv \emptyset$). از این‌رو معیار مشابهت BCF براساس جدول ۱، تابعی از ضریب BC بین یک جفت از کالاهای امتیازدهی شده در مشابهت محلی امتیازات یک جفت از کالاهاست.

زن و همکارانش سه نوع مختلف [۳۳، ۳۴، ۳۵] دیگری از پالایش مشارکتی براساس سامانه‌های توصیه‌گر برای محیط‌های گروهی مشارکتی ارائه نموده‌اند. به دلیل اینکه از معیارهای رایج نمی‌توان در این سناریو استفاده کرد، این است که آن‌ها نوعی از معیار مشابهت را ارائه نموده‌اند [۳۳، ۳۴] که براساس اطلاعات محتوایی جمع‌آوری شده از گروه‌های مشارکتی اقدام به معرفی معیار مشابهت جدیدی را کرده‌اند. مدل فضایی جریان کاری چهاربعدی به‌منظور محاسبه شباهت میان اعضای گروه مورد استفاده قرار می‌گیرد. همچنین زن و همکارانش در [۳۵] معیار دیگری را به‌منظور محاسبه شباهت بین همتایان برای سامانه‌های توصیه‌گر مبتنی بر دانش در محیط‌های P2P ارائه نموده‌اند. اطلاعات کاربری هر کاربر شامل خصیصه‌های عددی، دودویی و اسمی می‌باشد. این معیارها در حوزه بسیار خاص کاربرد دارند.

۳- معیار شباهت پیشنهادی برای پالایش مشارکتی

در اکثر سامانه‌های توصیه‌گر، بیشتر کاربران فقط به تعداد کمی از کالاها امتیاز می‌دهند. از این‌رو معیارهای مشابهت رایجی که در فصل قبلی تشریح گردید در مجموعه داده‌ای که امتیازات کاربران در آن به صورت پراکنده باشد از عملکرد پایینی برخوردار می‌باشند. در این بخش می‌خواهیم معیار مشابهتی را پیشنهاد دهیم که هم مناسب مجموعه داده‌های غیرپراکنده بوده و هم در مجموعه داده پراکنده از کارایی بالایی برخوردار است.

معیار مشابهت پیشنهادی ما از رویکردی احتمالاتی برای مدل‌سازی و فرموله‌بندی شباهت میان دو کاربر استفاده گردیده است. از این‌رو ابتدا رویکرد احتمالاتی معیار مشابهت پیشنهادی خود بیان کرده و سپس رویکرد خود را در مجموعه داده‌های دودویی استفاده می‌کنیم و در نهایت ایده اصلی معیار مشابهت پیشنهادی خود را به منظور استفاده در مجموعه داده‌های واقعی تعمیم می‌دهیم.

با فرض اینکه افراد جامعه نسبت به یک کالا خاص نظرات خود را در قالب نظرات مثبت و منفی (نظرات دودویی) اظهار می‌نمایند؛ جامعه به دو گروه آراء مثبت و آراء منفی افراز می‌گردد. در صورتیکه بخواهیم شباهت میان دو کاربر را محاسبه نماییم، آن دو کاربر می‌توانند هم‌عقیده (نظرات یکسان) بوده و یا عقاید متفاوتی (نظرات متفاوت) نسبت به یکدیگر داشته باشند.

از این‌رو به منظور محاسبه احتمال انتخاب دو کاربر u و v از میان کل کاربرانی که به آن کالای خاص امتیاز داده‌اند از طرق معادله (۳) قابل محاسبه خواهد بود:

$$P(u,v)=\begin{cases}\frac{\binom{m}{2}}{\binom{n}{2}} & \text{کاربر u و v با یکدیگر هم عقیده باشند}\\ \frac{\binom{n_1}{1}\times\binom{n_2}{1}}{\binom{n}{2}} & \text{در غیرانصورت}\end{cases} \quad (۳)$$

که براساس معادله (۳)، m تعداد کاربرانی است که نظرات مشابه‌ی نسبت به یکدیگر را به یک کالا دارند و n نیز بیانگر تعداد کل کاربرانی است که نظرات خود را نسبت به آن کالای خاص داده‌اند، می‌باشد. همچنین n_1 بیانگر تعداد کاربرانی است که نظرات مشابهی را نسبت به کاربر u و n_2 بیانگر تعداد کاربرانی است که نظرات مشابهی را نسبت به کاربر v، به یک کالا داده‌اند.

۳-۱- مثالی از معیار مشابهت پیشنهادی

یک مورد نظرسنجی را در نظر بگیرید که در آن افراد جامعه نسبت به یک کالا خاص نظرات خود را در قالب نظرات مثبت و منفی (نظرات دودویی) اظهار می‌نمایند؛ بنابراین بعد از جمع‌آوری نظرات، جامعه به دو گروه آراء مثبت و آراء منفی افراز می‌گردد. ایده اصلی برای محاسبه شباهت میان افراد بر مبنای اندازه جمعیت تشکیل‌دهنده این گروه‌ها و تقارن جامعه شکل‌گرفته استوار است.

اگر دو نفر را به صورت کاملاً تصادفی از میان جامعه شکل گرفته با شاکله فوق انتخاب نماییم، آن‌ها می‌توانند هم‌عقیده بوده (نظرات یکسان) و یا عقاید متفاوتی (نظرات متفاوت) نسبت به یکدیگر داشته باشند؛ بنابراین با در نظر گرفتن افراز صورت گرفته سه حالت زیر را می‌توان متصور بود:

۱. دو فرد با یکدیگر هم‌عقیده بوده و هردوی آن‌ها در گروه اقلیت قرار دارند.
۲. دو فرد با یکدیگر هم‌عقیده بوده و هردو آن‌ها در گروه اکثریت قرار دارند.
۳. دو فرد با یکدیگر هم‌عقیده نبوده و نظرات متفاوتی را نسبت بهم دارند.

جامعه‌ای را در نظر بگیرید که شامل ۲۰ نفر می‌گردد. در صورتیکه آرای دودویی جمع‌آوری شده برای کالای iام، جامعه موردنظر را به دو افراز ۱۸ و ۲ نفری (مثبت و منفی) تقسیم نماید، احتمال انتخاب دو نفر هم‌عقیده در گروه اقلیت (حالت ۱) بسیار کمتر از احتمال یافتن آن‌ها در گروه اکثریت (حالت ۲) می‌باشد؛ چرا که $\binom{18}{2} \gg \binom{2}{2}$.

همچنین احتمال انتخاب هم‌عقیده نبودن دو نفر، هنگامی‌که در گروه نامتقارن قرار بگیرند کمتر از احتمال یافتن آن‌ها در گروه متقارن می‌باشد؛ چرا که $\binom{18}{1}\times\binom{2}{1} < \binom{10}{1}\times\binom{10}{1}$.

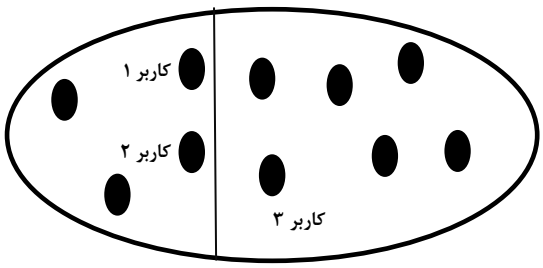
حال برآیند آنچه را که در تشریح داده شد را می‌توان به این صورت برداشت نمود که احتمال حالات ۳ از احتمال حالت ۲ کمتر و از احتمال حالت ۱ بیشتر می‌باشد؛ چرا که $\binom{18}{2} > \binom{10}{1}\times\binom{10}{1} > \binom{18}{1}\times\binom{2}{1} > \binom{2}{2}$.

۳-۲- ایده اصلی معیار مشابهت پیشنهادی

از این‌رو در این مقاله از رویکرد احتمالاتی برای مدل‌سازی معیار مشابهت میان دو کاربر استفاده نموده‌ایم. در واقع احتمال هم‌عقیده بودن و یا هم‌عقیده نبودن این دو نفر با توجه به عضویت آن‌ها در گروه‌های اکثریت و یا اقلیت به‌عنوان معیاری برای شباهت میان افراد قلمداد می‌گردد.

مطابق شکل ۱ از آنجایی‌که احتمال بیشتری (کمتری) برای پیدا کردن افراد هم‌عقیده در گروه‌های بزرگ‌تر (کوچکتر) وجود دارد، شباهت کمتری (بیشتری) را

برای آن دو نفر قائل هستیم. با در نظر گرفتن اینکه اندازه حداکثر احتمالات برابر با یک می‌باشد، بنابراین از اندازه مکمل آن برای محاسبه معیار شباهت میان افراد استفاده می‌کنیم. به‌عبارت دیگر معیار مشابهت پیشنهادی بیشترین تأثیر را برای حالتی دارد که اندازه جوامع آراء مثبت و منفی بیشترین فاصله را با یکدیگر داشته باشند



شکل ۱- هم عقیده بودن کاربر ۱ و ۲ و هم عقیده نبودن آنها با کاربر ۳ در کالای iام

از این‌رو به منظور محاسبه شباهت میان دو کاربر u و v از میان کل کاربرانی که به آن کالا iام امتیاز داده‌اند در صورتیکه با یکدیگر هم‌عقیده باشند و یا نظرات متفاوتی نسبت به یکدیگر داشته باشند، از طریق معادله (۴) قابل محاسبه می‌باشد.

$$S^{i}(u,v)=\begin{cases}1-\frac{\binom{m}{2}}{\binom{n}{2}} & \text{کاربر u و v با یکدیگر هم عقیده باشند}\\ 1-\frac{\binom{n_1}{1}\times\binom{n_2}{1}}{\binom{n}{2}} & \text{در غیرانصورت}\end{cases} \quad (۴)$$

که براساس معادله (۴)، m تعداد کاربرانی است که آرای یکسانی نسبت به یکدیگر را به یک کالا داشته باشند و n نیز بیانگر تعداد کل کاربرانی است که نظرات خود را نسبت به آن کالای خاص داده‌اند، می‌باشد. همچنین n_1 بیانگر تعداد کاربرانی است که آرای یکسانی را نسبت به کاربر u و n_2 بیانگر تعداد کاربرانی است که آرای یکسانی را نسبت به کاربر v، به یک کالا داده‌اند.

در نهایت برای بدست آوردن شباهت میان کاربران برای حالتی که کاربران به چندین کالای مختلف امتیاز داده باشند، می‌توانیم با محاسبه میانگین مجموع مشابهت به‌دست آمده به‌ازای هر کالا (i)، مشابهت میان دو کاربر در حالت امتیازات دودویی به دست آوریم. بنابراین از معادله (۵) به‌منظور محاسبه معیار پیشنهادی بین کاربر U و V استفاده می‌نماییم:

$$S(U,V)=\frac{\sum_{i=1}^{|i|}S^{i}(u,v)}{|i|} \quad (۵)$$

که براساس معادله (۵)، $S^{i}(u,v)$ معیار مشابهت پیشنهادی دو کاربر u و v به ازاء یک کالا می‌باشد و $S(U,V)$ بیانگر معیار شباهت دو کاربر u و v خواهد بود. همچنین $|i|$ بیانگر مجموع کل تعداد کالاهاست.

۳-۳- تعمیم ایده پیشنهادی

اکنون با تعمیم مثال خود در مجموعه داده‌های واقعی (مجموعه داده‌هایی که امتیازات به کالاها در محدوده ۵-۰ می‌باشند)، میزان مشابهت میان دو کاربر u و v را در مجموعه داده‌های واقعی و بر روی کالاهای مختلف محاسبه می‌نماییم:

۱) براساس امتیازات داده شده برای هر یک از کالاها، میانگین امتیازات داده شده محاسبه می‌گردد.

۲) براساس معادله (۶) کاربرانی که امتیاز آنها بالاتر و یا برابر میانگین امتیازات داده شده به کالای iام باشند، در جامعه مثبت (هم‌عقیده) و در غیر این صورت دیگر کاربران در جامعه منفی (عقاید متفاوت) افراز می‌گردند.

$$\begin{cases}P_{+}=\left\{u|u\in U\,,\,r_{u\,,\,I_i}\geq mean(I_i)\right\}\\ P_{-}=\left\{u|u\in U\,,\,r_{u\,,\,I_i}<mean(I_i)\right\}\end{cases} \quad (۶)$$

که براساس معادله (۶)، $r_{u\,,\,I_i}$ بیانگر امتیاز داده شده توسط کاربر u به کالای iام بوده و $mean(I_i)$ نیز بیانگر میانگین تمامی امتیازات داده شده به کالای iام می‌باشد.

۳) اکنون برای محاسبه شباهت میان افراد ابتدا براساس امتیازات داده شده برای یک کالای خاص میانگین امتیازات محاسبه گردیده، سپس براساس معادلات (۶) جامعه دودویی تشکیل می‌گردد. در نهایت براساس ایده جامعه دودویی شباهت میان آن دو کاربر محاسبه می‌گردد.

۴-۳- بررسی مدل معیار مشابهت پیشنهادی

به‌منظور تبیین ایده پیشنهادی جدول ۲ را نشان‌دهنده امتیازات هر کاربر به یک کالا می‌باشد را از مقاله [۳۰] بهره‌برداری کرده تا بتوانیم به مقایسه معیار مشابهت پیشنهادی خود با دیگر معیارهای مشابهت بپردازیم. برای مثال در نظر گرفته شده این حالت را در نظر گرفته‌ایم که پنج کاربر به چهار کالا موجود در سامانه مطابق جدول زیر امتیاز مابین یک تا پنج را داده‌اند. از این‌رو امتیاز کاربرانی که به بعضی از کالاها امتیاز نداده‌اند را با عدد صفر نشان داده‌ایم که به‌منزله عدم‌تمایل به کالا مربوطه و یا نادیده گرفته شدن آن کالا توسط کاربر موردنظر می‌باشد. سپس با توجه به مثال آورده شده مشابهت بین کاربران را با توجه به معیار شباهت پیشنهادی محاسبه می‌نماییم. از این‌رو شکل ۲ نشان‌دهنده مشابهت بین کاربران بوده که با توجه به جدول ۲ مورد محاسبه قرار گرفته شده است. به دلیل اینکه ماتریس حاصله برای تمامی معیارهای مشابهت متقارن می‌باشد، ما تنها به نمایش بخشی از آن اکتفا می‌نماییم.

از این‌رو براساس معیار مشابهت پیشنهادی که در قسمت قبل آن را تشریح نموده‌ایم، قادر خواهیم بود که ماتریس مشابهت بین کاربران را یافته و در پایان از آن به‌منظور پیش‌بینی امتیاز کاربران استفاده نماییم.

جدول ۲- مثالی از ماتریس امتیازات کاربر ـ کالا

	کالای ۱	کالای ۲	کالای ۳	کالای ۴
کاربر ۱	۴	۳	۵	۴
کاربر ۲	۵	۳	۰	۰
کاربر ۳	۴	۳	۳	۴
کاربر ۴	۲	۱	۰	۰
کاربر ۵	۴	۲	۰	۰

$$\begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \end{array}\begin{array}{c} \begin{array}{cccc} u_2u_3 & & u_4u_5 & \end{array} \\ \begin{bmatrix} 0/475 & 0.725 & 0.450 & 0.400 \\ & 0.475 & 0.600 & 0.550 \\ & & 0.450 & 0.400 \\ & & & 0.725 \end{bmatrix} \end{array}$$

شکل ۲- ماتریس معیار مشابهت محاسبه شده براساس جدول ۲

	u_2	u_3	u_4	u_5
u_1	0.707	0.0	0.707	0.707
u_2		1.0	1.0	1.0
u_3			1.0	1.0
u_4				1.0

الف) PC

	u_2	u_3	u_4	u_5
u_1	1.0	0.577	-0.447	0.707
u_2		0.577	-0.447	0.707
u_3			-0.447	0.707
u_4				0.707

ب) COS

	u_2	u_3	u_4	u_5
u_1	0.49	0.96	0.42	0.49
u_2		0.49	0.74	0.96
u_3			0.42	0.49
u_4				0.9

ج) JMSD

	u_2	u_3	u_4	u_5
u_1	0.5615	1.0077	0.5615	0.5615
u_2		0.6995	1.2	1.2
u_3			0.6995	0.6995
u_4				1.2

د) $BCD_{(cor)}$

	u_2	u_3	u_4	u_5
u_1	0.743	1.0	0.167	0.506
u_2		0.743	0.162	0.763
u_3			0.167	0.506
u_4				0.767

ه) PIP

	u_2	u_3	u_4	u_5
u_1	0.02089	0.0552	0.00475	0.0244
u_2		0.0183	0.00464	0.03561
u_3			0.00636	0.02500
u_4				0.01531

و) NHSM

شکل ۳- معیارهای مشابهت بدست آمده تمامی معیارهای بحث شده براساس جدول ۲

	i_1	i_2	i_3	i_4
u_1	4.75	3.25	4	4
u_2	4.5	3.5	-	-
u_3	4.25	2.75	3.5	3.5
u_4	2.5	0.5	-	-
u_5	3.75	2.25	-	-

الف) PC

	i_1	i_2	i_3	i_4
u_1	5	3	3.5	4.5
u_2	5	3	-	4.25
u_3	4.5	2.5	4.5	3.5
u_4	2.5	0.5	-	1.75
u_5	4	2	-	3.25

ب) CPC

	i_1	i_2	i_3	i_4
u_1	4.66	3.33	3.5	4.5
u_2	4.5	3.5	-	-
u_3	4.5	2.5	4.5	3.5
u_4	2.31	0.68	-	-
u_5	3.83	2.16	-	-

ج) JMSD

	i_1	i_2	i_3	i_4
u_1	4.24	3.24	4.27	4.24
u_2	4.27	3.27	-	-
u_3	3.77	2.77	3.66	3.77
u_4	1.77	0.77	-	-
u_5	3.27	2.27	-	-

د) $BCD_{(cor)}$

	i_1	i_2	i_3	i_4
u_1	4.71	3.28	3.5	4.5
u_2	4.5	3.5	-	-
u_3	4.5	2.5	4.5	3.5
u_4	2.25	0.75	-	-
u_5	3.8	2.19	-	-

ه) PIP

	i_1	i_2	i_3	i_4
u_1	4.65	3.34	3.5	4.5
u_2	4.83	3.16	-	-
u_3	4.5	2.5	4.5	3.5
u_4	2.35	0.64	-	-
u_5	3.5	2.5	-	-

و) NHSM

	i_1	i_2	i_3	i_4
u_1	4.23	3.23	4.29	4.23
u_2	4.25	3.25	-	-
u_3	3.76	2.76	3.70	3.76
u_4	1.75	0.75	-	-
u_5	3.25	2.25	-	-

ز) معیار شباهت پیشنهادی

شکل۴- پیش‌بینی صورت گرفته با استفاده از معیارهای مشابهت متفاوت

در نهایت با محاسبه مشابهت بین کاربران با استفاده از دیگر معیارهای مشابهت مطابق با آنچه در جدول ۱ بیان شد، می‌پردازیم. شکل ۳ نشان‌دهنده نتایج حاصله از اجرای معیار مشابهت بحث شده بر روی جدول ۲ می‌باشد. به دلیل اینکه ماتریس حاصله برای تمامی معیارهای مشابهت متقارن می‌باشد، ما تنها به نمایش بخشی از آن اکتفا می‌نماییم.

از این‌رو براساس معیار مشابهت به دست آمده می‌توانیم عمده نقایص آن‌ها را بیان نماییم.

۱. قسمت (الف) شکل ۳ بیانگر معیار مشابهت مابین کاربران براساس معیار مشابهت پیرسون می‌باشد. براساس جدول ۲ می‌توانیم به‌راحتی مشاهده نماییم که کاربران ۱ و ۳ بیشترین میزان مشابهت را به یکدیگر دارا می‌باشند. بردار امتیازی آن‌ها برای کاربر ۱ (۴، ۳، ۵، ۴) و برای کاربر ۳ (۴، ۳، ۳، ۴) می‌باشد. با این حال همان‌گونه که در قسمت الف شکل ۳ مشاهده می‌نمایید، میزان مشابهت میان این دو کاربر، صفر در نظر گرفته است در عین اینکه بیشترین مشابهت را به یکدیگر دارا می‌باشند ولی دارای کمترین مقدار عددی

در معیار PC می‌باشند. همچنین با نگاهی به شکل ۳ می‌توان به این نکته پی برد که دیگر معیارهای مشابهت از این مشکل رنج نمی‌برند.

حال با توجه محاسبات صورت گرفته با استفاده از معیار مشابهت پیشنهادی در شکل ۲ می‌توان ملاحظه نمود که مقدار عددی مشابهت ما بین کاربر ۱ و ۳ برابر با ۰.۷۲۵ می‌باشد؛ که بالاترین میزان مشابهت مابین کاربران می‌باشد.

۲. قسمت (ب) شکل ۳ بیانگر معیار مشابهت ما بین کاربران براساس معیار مشابهت CPC می‌باشد. براساس جدول ۲ می‌توانیم به راحتی مشاهده نماییم که کاربران ۱ و ۲ از میزان مشابهت کمتری نسبت به یکدیگر برخوردارند، به این دلیل که بردار امتیازی برای کاربر ۱ (۴، ۳، ۵، ۴) و برای کاربر ۲ (۵، ۳، ۰، ۰) می‌باشد. با این حال همان‌گونه که در قسمت (ب) شکل ۳ مشاهده می‌نمایید، میزان مشابهت میان این دو کاربر، یک در نظر گرفته شده است. همچنین براساس آنچه در بالا بیان شد، میزان مشابهت بین کاربران ۱ و ۳ بیشترین مقدار عددی می‌بایست باشد که این موضوع نیز در معیار CPC نادیده گرفته شده است.

حال با توجه به محاسبات صورت گرفته با استفاده از معیار مشابهت پیشنهادی در شکل ۲ می‌توان ملاحظه نمود که مقدار عددی مشابهت مابین کاربران ۱ و ۲ برابر با ۰.۴۷۵ می‌باشد؛ که تقریباً پایین‌ترین میزان مشابهت مابین کاربران می‌باشد. همچنین با نگاهی به شکل ۳ می‌توان به این نکته پی برد که دیگر معیارهای مشابهت از این مشکل رنج نمی‌برند.

۳. قسمت (ه) و (و) شکل ۳ به ترتیب معیار مشابهت میان کاربران را براساس معیار مشابهت PIP و NHSM می‌باشد. براساس جدول ۲ می‌توان این برآورد را داشت که مشابهت که مابین کاربران ۳ و ۴ باید اندکی از مشابهت مابین کاربران ۲ و ۴ کمتر باشد، به دلیل اینکه بردار امتیازات مابین کاربران ۲ و ۴ نزدیکی بیشتری نسبت به یکدیگر را دارا می‌باشند. با این حال این موضوع در معیار PIP و NHSM دیده نمی‌شود، در حالی‌که در دیگر معیارهایی از قبیل JMSD و $BCD_{(cor)}$ این مسئله رعایت شده است.

حال با توجه به شکل ۲ که براساس معیار مشابهت پیشنهادی نشان داده شده است، می‌توان ملاحظه نمود که مقدار عددی مشابهت میان کاربران ۲ و ۴ از مقدار عددی مشابهت میان کاربران ۴ و ۳ بالاتر می‌باشد. همچنین با توجه به آنچه در موارد بالا به‌طور کامل تشریح شد، تنها معیار مشابهت JMSD توانسته است از دیگر مشکلات سایر معیارهای مشابهت رنج نبرده و عملکرد مناسبی را داشته باشد. با این حال در محاسبه پیش‌بینی امتیازات معیار مشابهت پیشنهادی عملکرد به‌مراتب بهتری نسبت به این الگوریتم داراست.

بعد از اینکه توانستیم ماتریس تمامی معیارهای مشابهت را محاسبه نماییم، قادر خواهیم بود که با استفاده از معادله (۱)، ماتریس امتیازات پیش‌بینی شده را نیز محاسبه نماییم. بر این اساس شکل ۴ بیانگر ماتریس امتیازات پیش‌بینی‌شده با استفاده از دو همسایه توسط هرکدام از معیارهای مشابهت بحث شده در جدول ۲ می باشد.

۴- آزمایش‌ها و ارزیابی روش پیشنهادی

به‌منظور ارزیابی عملکرد معیار پیشنهادی خود، ابتدا انواع مختلفی از معیارهای مشابهت را براساس پالایش مشارکتی مبتنی بر کاربران را اجرا نموده و در ادامه نتایج به دست آمده از معیار مشابهت خود را با آن‌ها مقایسه نموده‌ایم. از همین‌رو از معیارهای مشابهت رایجی از قبیل PC و CPC، معیار PIP (طراحی شده به‌منظور حل مشکل شروع سرد)، JMSD (ترکیبی از جاکارد و MSD)، و NHSM (بهبود یافته معیار PIP) و دو شکل مختلف BCD را به‌منظور مقایسه عملکرد الگوریتم خود با دیگر الگوریتم‌های سنتی و مدرن در نظر گرفته‌ایم.

۴-۱- نمایش مجموعه داده‌ها

از بین تمامی مجموعه داده‌ای معتبر منتشر شده، ما از مجموعه داده‌ای پرکاربرد MovieLens و Netflix در آزمایش‌های خود مورد استفاده قرار داده‌ایم. شرح مختصری از این مجموعه داده که شامل تعداد کاربران، تعداد کالاها، تعداد کل امتیازات داده شده به کالاها توسط کاربران، شاخص پراکندگی و محدوده متیازات می‌گردد، در جدول ۳ آمده است. سطح پراکندگی را با استفاده از شاخص تراکم (K) که درصد همه امتیازات ممکن در دسترس در یک مجموعه داده‌ای می‌باشد را نشان داده‌ایم.

جدول ۳- خلاصه‌ای از مجموعه داده‌ای مورد استفاده در آزمایش‌ها

مجموعه داده	نام اختصار	تعداد کاربران (M)	تعداد کالاها (N)	تعداد امتیازات (R)	K $(\frac{R \times 100}{N \times M})$	محدوده امتیازات
MovieLens	ML_{1M}	۶۰۴۰	۳۹۵۲	۱۰۰۰۲۹۲	۴.۱۹	{۱،۲،۳،۴،۵}
NetFlix	NF_{100M}	۲۶۴۹۴۲۹	۱۷۷۷۰	۱۰۰۴۸۰۵۰۷	۲.۱۳	{۱،۲،۳،۴،۵}

به‌منظور نمایش اثربخشی معیار مشابهت پیشنهادی ارائه شده در داده‌های پراکنده، دو زیرمجموعه در سطوح مختلف پراکندگی را با استفاده از حذف تصادفی امتیازات از مجموعه داده‌ای MovieLens و Netflix به دست آمده است. ویژگی‌های این زیرمجموعه‌ها را در جدول ۴ به‌طور خلاصه بیان‌شده است.

جدول ۴- آماره‌های زیرمجموعه‌های پراکندگی

مجموعه داده	زیرمجموعه داده	کاربران (M)	کالاها (N)	تعداد امتیازات (R)	K $(\frac{R \times 100}{N \times M})$	$(\frac{R}{M})$	$(\frac{R}{N})$
MovieLens	ML_1	۶۰۴۰	۳۷۰۶	۴۰۹۵۷	۰.۱۸	۶.۸	۱۱.۱
NetFlix	NF_1	۲۶۴۹۴۲۹	۱۷۰۷۰	۹۰۴۵۱۵۰۶	۰.۲	۳۴.۱۳	۵۲۹۸.۸

۴-۲- معیارهای ارزیابی

در این مقاله به‌منظور اعتبارسنجی معیار پیشنهادی، مجموعه داده‌های موجود را به صورت تصادفی به دو زیرمجموعه آموزش (برای ساخت مدل) و آزمونه (برای تخمین دقت) تقسیم می‌نماییم که به روش Hold-Out Cross Validation موسوم است. زیرمجموعه آموزش شامل ۸۰٪ کل امتیازات و زیرمجموعه آزمونه شامل ۲۰٪ مابقی امتیازات می‌باشد. سپس مدل موردنظر با استفاده از داده‌های آموزشی، آموزش داده شده و نتیجه آن با استفاده از داده‌های آزمونه اعتبارسنجی می‌شود [۳۶].

معیارهای دقت پیش‌بینی: این معیارها دقت کمی مقدار امتیاز پیش‌بینی شده برای هر کالا را ارزیابی می‌نمایند. دو معیار اصلی مورد استفاده در روش ارزیابی آماری عبارت‌اند از: MAE و RMSE. در محاسبات لازم است برای مشخص کردن مقادیر این دو معیار، از دو بردار r و $\bar{r}$ استفاده کرد که به ترتیب دربردارنده امتیازت داده شده توسط کاربر و امتیازات پیش‌بینی شده توسط سامانه توصیه‌گر هستند و طول بردارها را نیز با Max نشان داده می‌شود. بدین‌ترتیب پارامترهای ارزیابی مذکور با روابط (۷) و (۸) محاسبه می‌شوند.

$$MAE = \frac{\sum_{i=1}^{Max} |r_i - \bar{r}_i|}{Max} \quad (۷)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{Max} (r_i - \bar{r}_i)^2}{Max}} \quad (۸)$$

هر چه مقادیر MAE و RMSE کوچک‌تر باشند، به این معناست که صحت پیش‌بینی‌های سامانه توصیه‌گر موردنظر بیشتر است.

معیارهای صحت دسته‌بندی: این معیارها عملکرد کیفی یک سامانه توصیه‌گر را ارزیابی می‌نمایند. اکثر سامانه‌های توصیه‌گر بجای پیش‌بینی کالاها، فهرستی از کالاها (L_r) را برای کاربر هدف مهیا می‌نمایند. یکی از دو معیار معروف برای ارزیابی کیفی یک سامانه توصیه‌گر معیار precison بوده که اختلاف مابین وابستگی کالاهای موجود در لیست L_r را براساس معادله (۹) محاسبه نموده و معیار دیگر recall بوده که اختلاف مابین وابستگی کل کالاها با کالاهای موجود در لیست L_r را براساس معادله (۱۰) محاسبه می‌نماید. از همین‌رو یک لیست کالاهای مرتبط L_{rev} برای یک کاربر، مجموعه کالاهایی است که کاربر هدف به آن‌ها بالاترین امتیاز (یعنی امتیاز بزرگ‌تر از ۴) را داده است.

$$precison = \frac{|L_r \cap L_{rev}|}{|L_r|} \quad (۹)$$

$$recall = \frac{|L_r \cap L_{rev}|}{|L_{rev}|} \quad (۱۰)$$

با این حال همواره موازنه‌ای مابین این دو معیار در نظر گرفته می‌شود. از این‌رو کاهش تعداد کالاها در L_r موجب افزایش recall و در عین حال باعث کاهش precison می‌گردد. بنابراین از معیاری در آزمایش‌های خود استفاده نموده‌ایم که هر دوی این معیارهای را ترکیب کرده که با نام F_β شناخته می‌شود و مطابق معادله (۱۱) محاسبه می‌شود. به‌طوری‌که پارامتر $\beta \in [0,1]$ تأثیر نسبی هر دو سنجه را تعیین می‌کند (معمولاً مقدار $\beta = 1$ مورد استفاده قرار می‌گیرد).

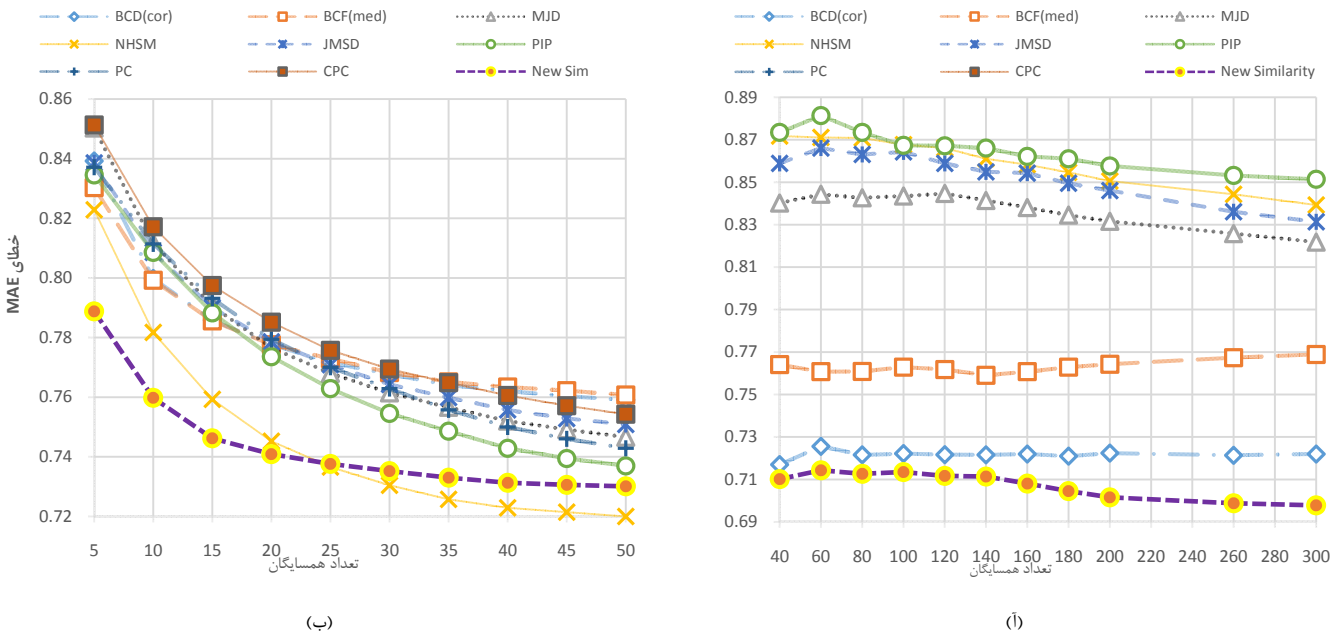
$$F_\beta = \frac{(1+\beta^2)\times precision \times recall}{\beta^2 \times (precison + recall)} \quad (۱۱)$$

۴-۳- نتایج و تحلیل آزمایش‌ها

۴-۳-۱- نتایج خطای MAE

شکل ۵ میزان خطای MAE محاسبه شده بعد از اجرای معیار مشابهت مختلف بر روی زیرمجموعه ML_1 و مجموعه ML_{1M} می‌باشد. با توجه به اینکه معیارهای مشابهت $BCF_{(cor)}$ و $BCF_{(med)}$ از تمامی اطلاعات امتیازات به‌منظور تعیین همسایگان کاربر هدف استفاده می‌نمایند، اما در هر حال نتوانسته‌اند نسبت به معیار مشابهت پیشنهادی خطای کمتری را داشته باشند و از این‌رو معیار مشابهت پیشنهادی توانسته است خطای کمتری نسبت به آن دو معیار داشته باشد. حال با توجه به اینکه دیگر معیار مشابهت سنتی تنها از اطلاعاتی استفاده می‌نمایند که توسط هر دو کاربر امتیاز داده شده باشند، از این‌رو در زیرمجموعه داده ML_1 که بسیار پراکنده بوده و تنها ۴۰۹۵۷ امتیاز در این مجموعه داده موجود می‌باشد، این امر میسر نبوده و نمی‌توان همسایگان دقیق کاربر هدف را پی برد. به همین ترتیب این‌چنین معیارهایی از خطایی بالاتری نسبت به دیگر معیارها برخوردار می‌باشند.

براساس قسمت (آ) شکل ۵ عملکرد معیار مشابهت پیشنهادی در داده‌های پراکنده نسبت به تمامی معیار مشابهتی که اخیراً ارائه شده‌اند از خطای کمتری برخوردار است و دیگر معیارهای مشابهت با تعداد همسایگان مختلف، دارای خطای MAE بیشتر از ۰.۶۹۷۸ می‌باشند. از بین تمامی معیار مشابهت مقایسه شده، معیار مشابهت $BCF_{(cor)}$ رقابت نزدیک‌تری را با معیار مشابهت پیشنهادی داشته و از این‌رو دارای خطای MAE بیشتر از ۰.۷۱۷۰ می‌باشد. همچنین در بازه تمامی همسایگان، معیار مشابهت پیشنهادی دارای متوسط خطای ۰.۷۰۷۷ می‌باشد در حالی‌که متوسط خطای نزدیکترین رقیب معیار مشابهت پیشنهادی ($BCF_{(cor)}$) برابر ۰.۷۲۷۱ است که نشان‌دهنده این است که معیار مشابهت پیشنهادی بیشتر از ۲.۷٪ خطای کمتری برخوردار است.



(ب) (آ)

شکل ۵- خطای MAE در مقیاس تعداد همسایگان (آ) در زیرمجموعه ML_1 و (ب) در مجموعه ML_{1M}

براساس قسمت (ب) شکل ۵ نیز در تعداد همسایگی ۵-۲۰، معیار مشابهت پیشنهادی در داده‌های غیرپراکنده نسبت به تمامی معیارهای مشابهتی، از خطای به مراتب کمتری برخوردار بوده است، اما این روند در تعداد همسایگی ۲۵-۵۰ امکان نبوده و معیار NHSM از خطای کمتری نسبت به معیار مشابهت پیشنهادی برخوردار است. همچنین در بازه تمامی همسایگان، معیار مشابهت پیشنهادی دارای متوسط خطای ۰.۷۴۳۴ می‌باشد در حالی‌که متوسط خطای معیار مشابهت NHSM برابر ۰.۷۴۶۶ است که نسبت به معیار مشابهت پیشنهادی تنها از ۰.۴٪ خطای کمتری برخوردار است. با این حال معیار مشابهت پیشنهادی از دیگر معیارهای سنتی ارائه شده عملکرد بهتری داشته و این خود دلیلی بر کارایی بهتر معیار مشابهت پیشنهادی می‌باشد.

شکل ۶ میزان خطای MAE محاسبه شده بعد از اجرای معیار مشابهت مختلف بر روی زیرمجموعه NF_1 در داده‌های پراکنده و مجموعه NF_{100M} در داده‌های غیرپراکنده می‌باشد. با توجه به شکل ۶ می‌توان دریافت که در مجموعه داده Netflix به علت دارا بودن امتیازات زیاد معیار مشابهت پیشنهادی توانسته است هم در داده‌های پراکنده و هم در داده‌های غیرپراکنده عملکرد مطلوبی را از خود نشان دهد.

براساس قسمت (آ) شکل ۶، خطای معیار مشابهت پیشنهادی در داده‌های پراکنده (بجز معیار مشابهت $BCF_{(cor)}$ و $BCF_{(med)}$ و در بازه همسایگی ۴۰-۶۰) نسبت به تمامی معیار مشابهتی که اخیراً ارائه شده‌اند، عملکرد مطلوبتری را داشته و همانند دیگر معیار مشابهت دارای خطای ثابت ۰.۸۱۹۵ می‌باشد. در حالی‌که معیار مشابهت BCF روند صعودی داشته و تنها در همسایگی ۴۰-۶۰ توانسته‌اند عملکرد بهتری را داشته باشند و در مابقی همسایگان و با توجه به روند صعودی خطای خود، در همسایگی بالا عملکرد مثبتی از خود نشان نخواهند داد.

قسمت (ب) شکل ۶ نیز نشان‌دهنده مقایسه خطای معیار مشابهت پیشنهادی نسبت به دیگر معیارهای مشابهت بررسی شده می‌باشد که از این‌رو نشان‌دهنده عملکرد به‌مراتب بهتر معیار مشابهت پیشنهادی نسبت به دیگر رقبای خود می‌باشد. همچنین در بازه تمامی همسایگان، معیار مشابهت پیشنهادی دارای متوسط خطای ۰.۷۴۲۵ می‌باشد در حالی‌که متوسط خطای نزدیکترین رقیب معیار مشابهت پیشنهادی ($BCF_{(cor)}$) برابر ۰.۷۴۹۴ است که نشان‌دهنده این است که معیار مشابهت پیشنهادی در حدود ۰.۱٪ خطای کمتری دارا است.

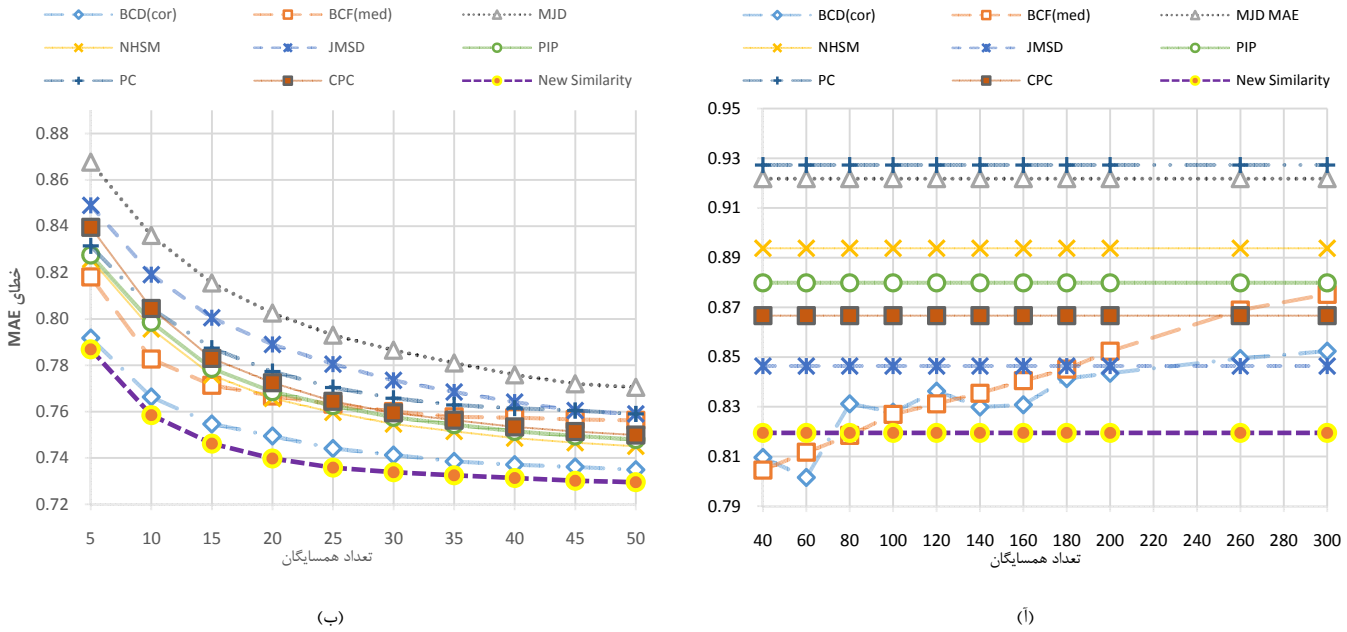
با توجه به شکل‌های ۵ و ۶ که نشان‌دهنده مقایسه معیار مشابهت پیشنهادی با دیگر معیارهای مشابهت سنتی و مدرن براساس خطای MAE می‌باشد، می‌توان به این نتیجه پی برد که معیار مشابهت در داده‌های پراکنده و غیرپراکنده در صورتیکه امتیازات کافی دردسترس باشد عملکرد بهتری را داراست و در غیر این صورت در تعداد همسایگی بالاتر می‌توان به نتایج مطلوبتری دست یافت.

مطابق قسمت (آ) شکل ۵ و ۶ می‌توان مشاهده نمود که خطای بدست آمده تقریباً خطی بوده و با توجه به تعداد امتیازات کم موجود، رابطه همسایگی میان کاربران به ازاء کالاهای متفاوت یکسان بوده و افزایش تعداد همسایگان تغییری در وضعیت آنها ایجاد نمی‌نماید.

۴-۳-۲- نتایج خطای RMSE

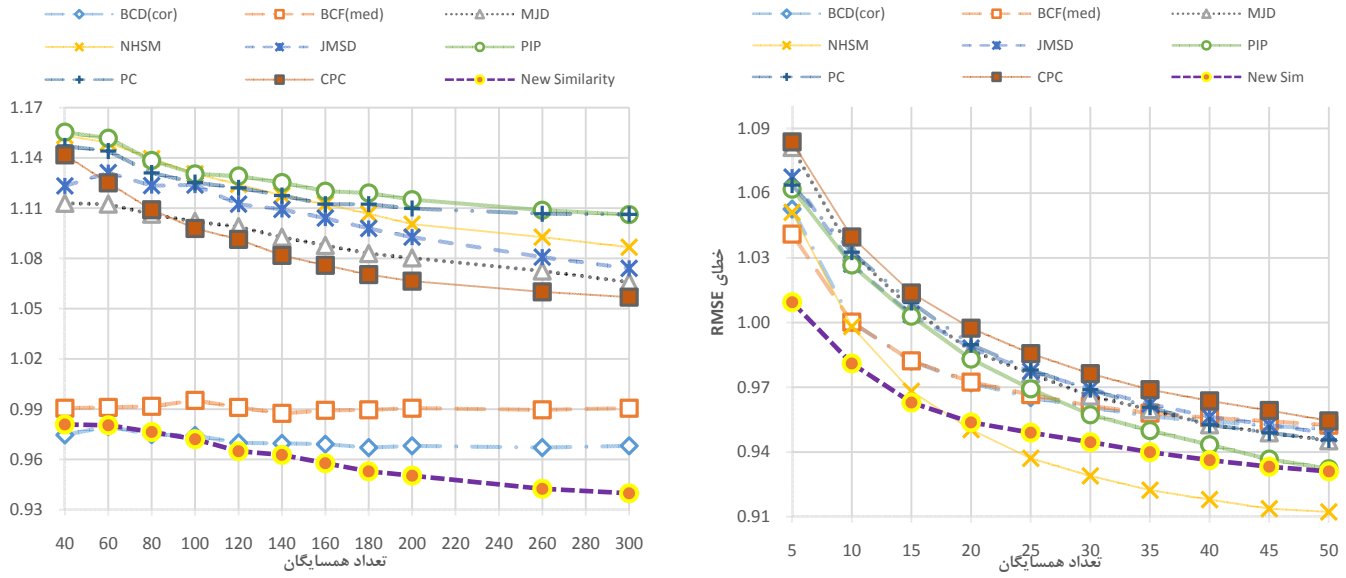
شکل ۷ نیز بیانگر میزان خطای RMSE محاسبه شده به‌ازای تعداد همسایگی‌های متفاوت بر روی زیرمجموعه ML_1 و مجموعه ML_{1M} می‌باشد.

براساس قسمت (آ) شکل ۷ معیار مشابهت پیشنهادی نسبت به‌تمامی معیار مشابهتی که اخیراً ارائه شده‌اند از خطای قابل ملاحظه کمتری برخوردار است و دیگر معیارهای مشابهت با تعداد همسایگی مختلف دارای خطای RMSE بیشتر از ۰.۹۶۷۱ می‌باشد. از بین تمامی معیار مشابهت مقایسه شد هم عیار مشابهت $BCF_{(cor)}$ رقابت نزدیک‌تری را نسبت به دیگر معیارهای مشابهت داشته و با این‌حال معیار مشابهت پیشنهادی بر تمامی معیارهای مشابهت برتری داشته و دارای خطای RMSE کمتر از ۰.۹۳۹۸ می‌باشد که نسبت به $BCF_{(cor)}$ حدود ۰.۳٪ کاهش خطا را داراست. همچنین در بازه تمامی همسایگان، معیار مشابهت پیشنهادی دارای متوسط خطای ۰.۹۶۱۹ می‌باشد در حالی‌که متوسط خطای نزدیکترین رقیب معیار مشابهت پیشنهادی ($BCF_{(cor)}$) برابر ۰.۹۷۱۱ است که نشان‌دهنده این است که معیار مشابهت پیشنهادی حدود ۰.۱٪ خطای کمتری برخوردار است.



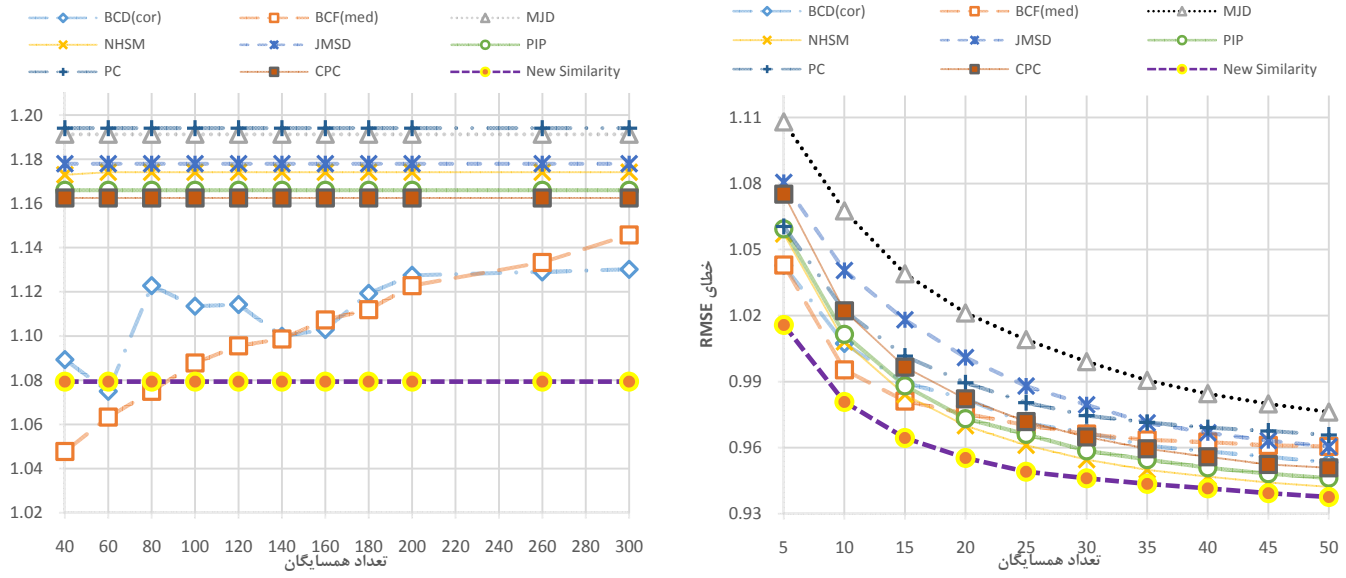
(آ) (ب)

شکل ۶- خطای MAE در مقیاس تعداد همسایگان (آ) در زیرمجموعه NF_1 و (ب) در مجموعه NF_{100M}



(ب) (آ)

شکل ۷- خطای RMSE در مقیاس تعداد همسایگان (آ) در زیرمجموعه ML_1 و (ب) در مجموعه ML_{1M}



(ب) (آ)

شکل ۸- خطای RMSE در مقیاس تعداد همسایگان (آ) در زیرمجموعه NF_1 و (ب) در مجموعه NF_{100M}

براساس قسمت (ب) شکل ۷ نیز در تعداد همسایگی ۵-۲۰، معیار مشابهت پیشنهادی در داده‌های غیرپراکنده نسبت به تمامی معیارهای مشابهتی، از خطای به‌مراتب کمتری برخوردار بوده است، اما این روند در تعداد همسایگی ۲۰-۵۰ امکان نبوده و معیار NHSM از خطای کمتری نسبت به معیار مشابهت پیشنهادی برخوردار است.

همچنین در بازه تمامی همسایگان، معیار مشابهت پیشنهادی دارای متوسط خطای ۰.۹۵۴۱ می‌باشد در حالی‌که متوسط خطای معیار مشابهت NHSM برابر ۰.۹۵۰۰ است که نسبت به معیار مشابهت پیشنهادی تنها از ۰.۴٪ خطای کمتری برخوردار است. با این حال معیار مشابهت پیشنهادی از دیگر معیارهای سنتی ارائه‌شده عملکرد بهتری داشته و این خود دلیلی بر کارایی بهتر معیار مشابهت پیشنهادی می‌باشد.

شکل ۸ نیز بیانگر میزان خطای RMSE محاسبه شده به‌ازای تعداد همسایگی‌های متفاوت بر روی زیرمجموعه NF_1 در داده‌های پراکنده و مجموعه NF_{100M} در داده‌های غیرپراکنده می‌باشد. با توجه به شکل می‌توان دریافت که در مجموعه داده Netflix به علت دارا بودن امتیازات زیاد معیار مشابهت پیشنهادی توانسته است تاحدودی در داده‌های پراکنده و غیرپراکنده عملکرد مطلوبی را از خود نشان دهد.

براساس قسمت (آ) شکل ۸، خطای معیار مشابهت پیشنهادی در داده‌های پراکنده (بجز معیار مشابهت $BCF_{(cor)}$ و $BCF_{(med)}$ و در بازه همسایگی ۴۰-۸۰)

نسبت به تمامی معیار مشابهتی که اخیراً ارائه شده‌اند، عملکرد مطلوبتری را داشته و همانند دیگر معیار مشابهت دارای خطای ثابت ۱.۰۷۹۴ می‌باشد. در حالی‌که معیار مشابهت BCF روند صعودی داشته و تنها در همسایگی ۴۰-۸۰ توانسته‌اند عملکرد بهتری را داشته باشند و در مابقی همسایگان و باتوجه به روند صعودی خطای خود، در همسایگی بالا عملکرد مثبتی از خود نشان نخواهند داد.

قسمت (ب) شکل ۸ نیز نشان‌دهنده مقایسه خطای معیار مشابهت پیشنهادی نسبت به دیگر معیارهای مشابهت بررسی شده می باشد که از این‌رو نشان‌دهنده عملکرد به مراتب بهتر معیار مشابهت پیشنهادی نسبت به دیگر رقبای خود می‌باشد. همچنین در بازه تمامی همسایگان، معیار مشابهت پیشنهادی دارای متوسط خطای ۰.۹۵۷۴ می‌باشد در حالی‌که متوسط خطای نزدیکترین رقیب معیار مشابهت پیشنهادی ($BCF_{(cor)}$) برابر ۰.۹۷۷۹ است که نشان‌دهنده این است که معیار مشابهت پیشنهادی بیشتر از ۲٪ خطای کمتری داراست.

۴-۳-۳- نتایج معیار F_1

نتایج معیار F_1 حاصله از اجرای تمامی معیارهای شباهت بحث شده بر روی زیرمجموعه ML_1 و مجموعه ML_{1M} در شکل ۹ به نمایش درآمده است.

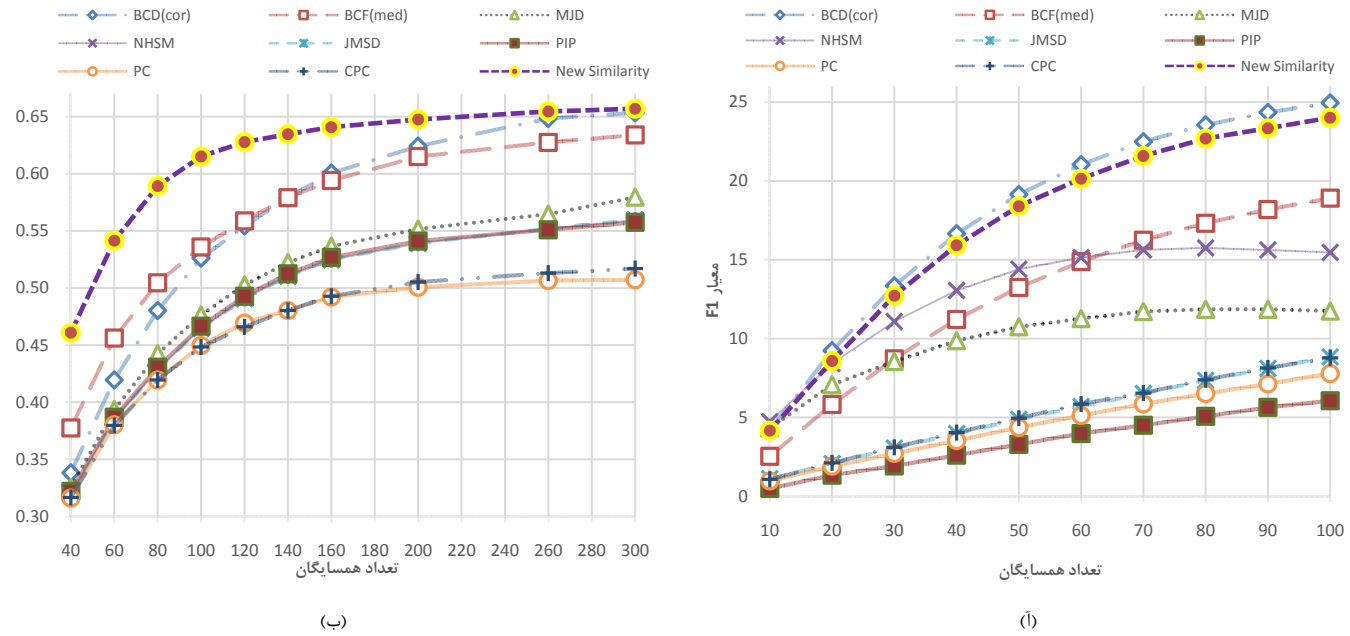
براساس قسمت (آ) شکل ۹ معیار مشابهت پیشنهادی عملکرد قابل توجهی نسبت به دیگر معیارهای مشابهت مبتنی بر همسایگان داراست. عملکرد معیار مشابهت پیشنهادی براساس معیار F_1 در بالاترین مقدار برابر ۰.۶۵۶۹ بوده در حالی‌که نزدیک‌ترین رقبای آن یعنی $BCD_{(cor)}$ و $BCD_{(med)}$ به ترتیب برابر ۰.۶۵۳۴ و ۰.۶۳۴۰است؛ یعنی اینکه دقت توصیه‌های صورت گرفته توسط معیار مشابهت پیشنهادی بیشتر از ۰.۵٪ نسبت به معیار $BCD_{(cor)}$ و بیشتر از ۳.۶٪ نسبت به معیار $BCD_{(med)}$ بالاتر است. به همین ترتی به معیار F_1 محاسبه شده برای نزدیکترین معیارهای مشابهت سنتی یعنی MJD کمتر از ۰.۵۷۹۵می‌باشدکه از این‌رو نشان‌دهنده این است که معیارهای مشابهت سنتی قادر به توصیه کالاهای مناسب به کاربر را نخواهند بود.

براساس قسمت (ب) شکل ۹ نتایج حاصله از معیار ارزیابی F_1 برای معیار مشابهت پیشنهادی اگر چه عملکردی ضعیف‌تر از معیار مشابهت $BCF_{(cor)}$ داشته ولی با توجه به روند صعودی نتایج حاصله از معیار مشابهت پیشنهادی نوید این می‌رود که در تعداد همسایگی بالا عملکرد معیار مشابهت پیشنهادی از دیگر رقبای خود بیشتر گردد. دیگر معیارهای مشابهت عملکرد مناسبی نداشته که بیانگر این است که دیگر معیار مشابهت در توصیه کالاهای مرتبط به کاربر هدف عملکرد پایینی را نسبت به دیگر معیارهای مشابهت دارا می‌باشند. مقدار عددی معیار ارزیابی F_1 در تعداد همسایگی ۵۰ تایی برای معیار مشابهت پیشنهادی در حدود ۰.۱۷۷۸ و برای معیار مشابهت $BCF_{(cor)}$ در حدود ۰.۱۸۹۲ بوده که نسبت به معیار پیشنهادی حدود ۷.۵٪ رشد را داراست.

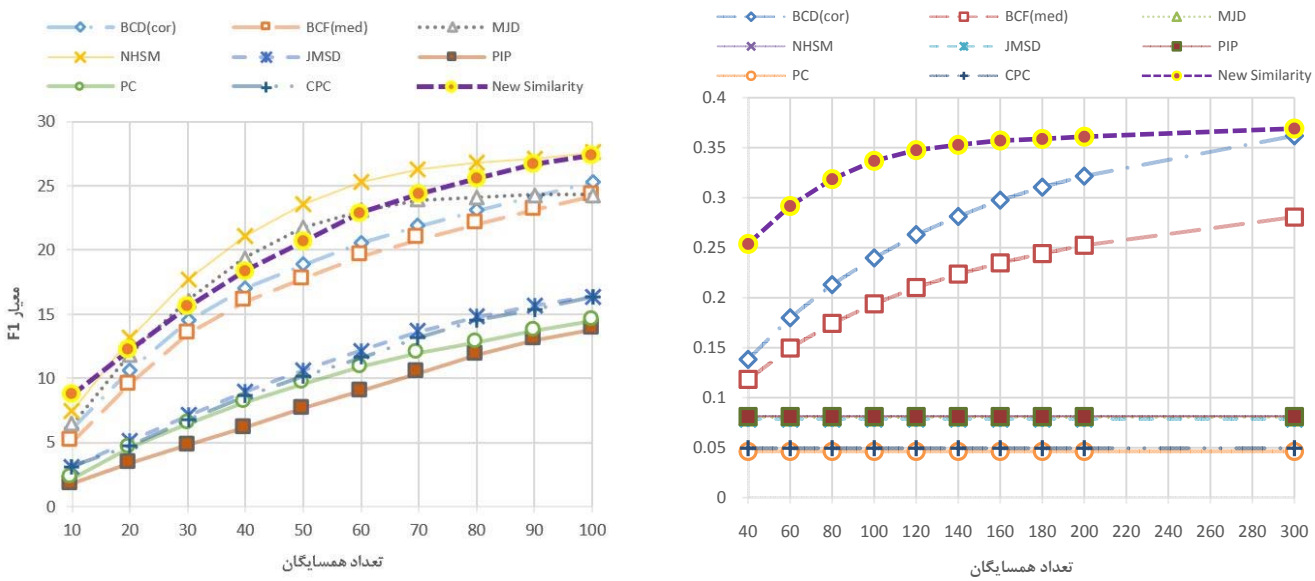
نتایج معیار F_1 حاصله از اجرای تمامی معیارهای شباهت بحث شده بر روی زیرمجموعه NF_1 و مجموعه NF_{100M} در شکل ۱۰ به نمایش درآمده است.

براساس قسمت (آ) شکل ۱۰ معیار مشابهت پیشنهادی عملکرد قابل توجهی نسبت به دیگر معیارهای مشابهت مبتنی بر همسایگان داراست. عملکرد معیار مشابهت پیشنهادی براساس معیار F_1 در بالاترین مقدار برابر ۰.۳۶۹۲ بوده در حالی‌که نزدیک‌ترین رقبای آن یعنی $BCD_{(cor)}$ و $BCD_{(med)}$ به ترتیب برابر ۰.۳۶۲۰ و ۰.۲۸۰۷ است؛ یعنی اینکه دقت توصیه‌های صورت گرفته توسط معیار مشابهت پیشنهادی حدود۲٪ نسبت به معیار $BCD_{(cor)}$ و بیشتر از ۳۱٪ نسبت به معیار $BCD_{(med)}$ بالاتر است. به همین ترتیب معیار F_1 محاسبه شده برای نزدیکترین معیارهای مشابهت سنتی یعنی MJD کمتر از ۰.۰۸۱۲ می‌باشد که از این‌رو نشان‌دهنده این است که معیارهای مشابهت سنتی قادر به توصیه کالاهای مناسب به کاربر را نخواهند بود.

براساس قسمت (ب) شکل ۱۰ نتایج حاصله از معیار ارزیابی F_1 برای معیار مشابهت پیشنهادی اگرچه عملکردی ضعیف‌تر از معیار مشابهت $BCF_{(cor)}$ داشته ولی با توجه به روند صعودی نتایج حاصله از معیار مشابهت پیشنهادی نوید این می‌رود که در تعداد همسایگی بالا عملکرد معیار مشابهت پیشنهادی از دیگر رقبای خود بیشتر گردد. دیگر معیارهای مشابهت عملکرد مناسبی نداشته که بیانگر این است که دیگر معیار مشابهت در توصیه کالاهای مرتبط به کاربر هدف عملکرد پایینی را نسبت به دیگر معیارهای مشابهت دارا می‌باشند.



شکل ۹- خطای F_1 در مقیاس تعداد همسایگان (آ) در زیرمجموعه ML_1و (ب) در مجموعه ML_{1M}



شکل ۱۰- خطای F_1 در مقیاس تعداد همسایگان (آ) در زیرمجموعه NF_1 و (ب) در مجموعه NF_{100M}

مقدار عددی معیار ارزیابی F_1 در تعداد همسایگی ۱۰۰ تایی برای معیار مشابهت پیشنهادی در حدود ۰.۲۷۳۶ و برای معیار مشابهت $BCF_{(cor)}$ در حدود ۰.۲۵۲۵ بوده که نسبت به معیار پیشنهادی بیشتر از ۸.۳٪ رشد را داراست.

۵- نتیجه‌گیری

در این مقاله از رویکرد احتمالاتی برای مدل‌سازی و فرمول‌بندی معیار شباهت میان دو کاربر استفاده شده است. بدین منظور ابتدا ایده خود را در حالت جامعه‌ای دودویی تبیین کرده سپس ایده خود را برای حالت چند امتیازی تعمیم داده‌ایم. در پایان این مقاله به این نتیجه رسیده‌ایم که معیار مشابهت پیشنهادی عملکرد بسیار خوبی نسبت به‌تمامی معیارهای مشابهت ارائه شده در داده‌های پراکنده داشته و در داده‌های غیرپراکنده نیز عملکرد قابل قبولی داشته و در آینده با بهبود روش‌ها خواهد توانست عملکرد بهتری نسبت به سایر معیارهای مشابهت داشته باشد.

هدف نهایی این مقاله ارائه یک معیار مشابهت نوآورانه‌ای است که بتواند ابتدا پیش‌بینی دقیق‌تری را به کاربر ارائه داده و نهایتاً از دقت و خطای کمتری برخوردار باشد. به‌علاوه برای گسترش تأثیرات نتایج این تحقیق، پیشنهادهایی وجود دارد که برای کارهای آتی می‌توان از آن‌ها استفاده نمود:

۱. می‌توان از دیگر روش‌های آماری همچون میانه، مد و غیره به‌منظور محاسبه مشابهت میان کاربران به‌ازای کالاهای مختلف استفاده نمود.

۲. با توجه به عملکرد مناسب این معیار در بهبود دقت سیستم‌های توصیه‌گر مبتنی بر پالایش مشارکتی، نتایج این تحقیق چشم‌انداز مناسبی جهت بهبود سایر تحقیقات در حوزه سامانه‌های توصیه‌گر ترکیبی نیز فراهم می‌آورد

مراجع

[1] A. J. Slywotzky, "The Age of Choiceboard," Harvard Business Review, vol. 78, no. 1, pp. 40-41, 2000.

[2] L. Terveen, and W. Hill, "Beyond recommender systems: Helping people help each other," HCI in the New Millennium, vol. 1, pp. 487-509, 2001.

[3] N. Landia, and S. Anand, "Personalised tag recommendation," Recommender Systems & the Social Web, New York, NY, USA, pp. 83-86, 2009.

[4] N. K. Ranjbar, and S. H. Alizadeh, "A fuzzy recommender system for forecasting customer segmentation by multi-variable fuzzy rule interpolation," in Proceedings of the Fuzzy Systems (IFSC), 2013 13th Iranian Conference on, 2013, pp. 1-5.

[5] D. Park, H. Kim, I. Choi, and J. Kim, "A literature review and classification of recommender systems research," Expert Systems with Applications, vol. 39, no. 11, pp. 10059-10072, 2012.

[6] L. Candillier, F. Meyer, and M. Boullé, "Comparing state-of-the-art collaborative filtering systems," Machine learning and data mining in pattern recognition, pp. 548-562, 2007.

[7] J. Schafer, D. Frankowski, and J. Herlocker, "Collaborative filtering recommender systems," The adaptive web, pp. 291-324, 2007.

[8] X. Sun, F. Kong, and S. Ye, "A comparison of several algorithms for collaborative filtering in startup stage," in Networking, Sensing and Control, 2005. Proceedings of the IEEE, 2005, pp. 25-28.

[9] X. Su, and T. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in artificial intelligence*, 2009, p. 4.

[10] C. Desrosiers, and G. Karypis, "A comprehensive survey of neighborhood-based recommendation methods," Recommender systems handbook, pp. 107-144, 2011.

[11] 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.

[12] L. Sheugh, and S. H.Alizadeh, "Merging similarity and trust based social networks to enhance the accuracy of trust-aware recommender systems," Journal of Computer & Robotics, vol. 8, no. 2, pp. 43-51, 2015.

[13] L. Sheugh, and S. H. Alizadeh, "Merging similarity and trust based social networks to enhance the accuracy of trust-aware recommender systems," Journal of Computer & Robotics, vol. 8, no. 2, pp. 43-51, 2015.

[14] J. Wang, A. D. Vries, and M. Reinders, "Unifying user-based and item-based collaborative filtering approaches by similarity fusion," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 501-508.

[15] L. Sheugh, and S. H.Alizadeh, "A note on pearson correlation coefficient as a metric of similarity in recommender system," in Proceedings of the AI & Robotics (IRANOPEN), 2015, 2015, pp. 1-6.

[16] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," Knowledge-based systems, vol. 46, pp. 109-132, 2013.

[17] Y. Koren, "Factor in the neighbors: Scalable and accurate collaborative filtering," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 4, no. 1, p. 1, 2010.

[18] Y. Koren, "Factor in the neighbors: Scalable and accurate collaborative filtering," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 4, no. 1, p. 1, 2010.

[19] K. Ali, and W. Van Stam, "TiVo: making show recommendations using a distributed collaborative filtering architecture," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 394-401.

[20] M. Ekstrand, J. Riedl, and J. Konstan, "Collaborative Filtering Recommender Systems," Foundations and Trends in Human-Computer Interaction, vol. 4, no. 2, pp. 81-173, 2011.

[21] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," IEEE Internet computing, vol. 7, no. 1, pp. 76-80, 2003.

[22] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in Proceedings of the 10th international conference on World Wide Web, 2001, pp. 285-295.

[23] G. G. CHOWDHURY, Introduction to modern information retrieval. London, England: Facet Publishing, 2010.

[24] U. Shardanand, and P. Maes, "Social information filtering: algorithms for automating "word of mouth," in Proceedings of the SIGCHI conference on Human factors in computing systems, 1995, pp. 210-217.

[25] 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.

[26] A. El-Saddik, H. Kim, and G. Jo, "Collaborative error-reflected models for cold-start recommender systems," Decision Support Systems, vol. 51, no. 3, pp. 519-531, 2011.

[27] J. Bobadilla, F. Ortega, and A. Hernando, "A collaborative filtering similarity measure based on singularities," Information Processing & Management, vol. 48, no. 2, pp. 204-217, 2012.

[28] J. Bobadilla, F. Ortega, and A. Hernando, "A collaborative filtering similarity measure based on singularities," Information Processing & Management, vol. 48, no. 2, pp. 204-217, 2012.

[29] J. Bobadilla, F. Ortega, A. Hernando, and J. Bernal, "A collaborative filtering approach to mitigate the new user cold start problem," Knowledge-Based Systems, vol. 26, pp. 225-238, 2012.

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

[31] B. K. Patra, R. Launonen, V. Ollikainen, and S. Nandi, "Exploiting Bhattacharyya similarity measure to diminish user cold-start problem in sparse data," in Proceedings of the International Conference on Discovery Science, 2014, pp. 252-263.

[32] B. K. Patra, R. Launonen, V. Ollikainen, and S. Nandi, "A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data," Knowledge-Based Systems, vol. 82, 2015.

[33] L. Zhen, G. Q. Huang, and Z. Jiang, "Collaborative filtering based on workflow space," Expert Systems with Applications, vol. 36, no. 4, pp. 7873-7881, 2009.

[34] L. Zhen, G. Q. Huang, and Z. Jiang, "Recommender system based on workflow," Decision Support Systems, vol. 48, no. 1, pp. 237-245, 2009.

[35] L. Zhen, Z. Jiang, and H. Song, "Distributed recommender for peer-to-peer knowledge sharing," Information Sciences, vol. 180, no. 18, pp. 3546-3561, 2010.

[36] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender

systems," ACM Transactions on Information Systems (TOIS), vol. 22, no. 1, pp. 5-53, 2004

مجتبی کاظمی مدرک کارشناسی مهندسی کامپیوتر را در سال ۱۳۹۰ از دانشگاه بین‌المللی امام خمینی (ره) اخذ نمود. او در سال ۱۳۹۵ دوره کارشناسی‌ارشد خود را در رشته مهندسی کامپیوتر در دانشگاه آزاد اسلامی قزوین به اتمام رسانید. علاقه‌مندی ایشان در زمینه داده‌کاوی، یادگیری ماشین و سامانه‌های توصیه‌گر می‌باشد.
آدرس پست‌الکترونیکی ایشان عبارت است از:
kazemi.mce@gmail.com

ساسان حسینعلی‌زاده مدرک کارشناسی مهندسی کامپیوتر را در سال ۱۳۸۲ از دانشگاه شیراز اخذ نمود. او در سال ۱۳۹۴ دوره کارشناسی‌ارشد خود را در رشته علوم کامپیوتر دانشگاه امیرکبیر به اتمام رسانید و در سال ۱۳۹۰ با درجه دکتری ریاضی کاربردی، گرایش کاربرد در کامپیوتر از همان دانشگاه فارغ التحصیل شد. هم‌اکنون او عضو هیات علمی دانشکده کامپیوتر و فناوری اطلاعات دانشگاه آزاد اسلامی قزوین، واحد قزوین می‌باشد. همچنین ایشان در زمینه‌های فرآیندهای تصادفی، یادگیری ماشین، سامانه‌های توصیه‌گر و شبکه‌های اجتماعی تحقیق می نماید.
آدرس پست‌الکترونیکی ایشان عبارت است از:
sasan.h.alizadeh@qiau.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۶/۰۳/۲۱
تاریخ اصلاح: ۱۳۹۶/۰۵/۰۷
تاریخ قبول شدن: ۱۳۹۶/۰۶/۰۱
نویسنده مرتبط: مجتبی کاظمی، دانشکده برق ـ رایانه و فناوری اطلاعات، دانشگاه آزاد اسلامی واحد قزوین، قزوین، ایران.

---

[1]Recommender Systems (RS)
[2]Social Information
[3]Followers
[4]Twits
[5]Content-Based Filtering (CBF)
[6]Demographic Filtering
[7]Collaborative Filtering (CF)
[8]Hybrid Filtering
[9]Sparsity
[10]Cold Start
[11]Co-Rated Item Set
[12]Pearson Correlation Coefficient (PCC)
[13]Adjusted Cosine (ACOS)
[14]Constrained PC (CPC)