مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۵ ، بهار ۱۳۹۶، ص ۱۹- ۸

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

مجلـه‌ی علمی پژوهش در علوم رایانه (سال دوم)
شماره ۵ / بهار ۱۳۹۶ / ص۱۹- ۸

# بهبود ضریب همبستگی پیرسون به منظور محاسبه شباهت میان کاربران، مبتنی بر رده بندی در سیستم‌های توصیه‌گر

**زکیه نیکدل[۱]، رضا قائمی[۲]**

[۱] گروه کامپیوتر، واحد نیشابور، دانشگاه آزاد اسلامی، نیشابور، ایران

[۲] گروه مهندسی کامپیوتر، واحد قوچان، دانشگاه آزاد اسلامی، قوچان، ایران

نام و نشانی ایمیل نویسنده مسئول:

**رضا قائمی**

***rezaghaemi73@gmail.com***

**چکیده**

هدف سیستم‌های پیشنهادگر یافتن آیتم‌های مورد علاقه از میان تعداد زیادی آیتم می‌باشد. ایده اصلی روش پالایش مشارکتی که از موفق‌ترین سیستم‌های پیشنهادی به شمار می‌رود، به این صورت است که اگر دو کاربر امتیازدهی‌های یکسان بر روی آیتم‌های مشترک داشته باشند، انگاه آن‌ها علاقه‌های یکسانی دارند. یکی از کلیدی‌ترین اجزا در سیستم‌های پیشنهادگر، بخش پیدا کردن همسایه‌های کاربر فعال می‌باشد که اگر به‌درستی انتخاب شود، می‌تواند صحت پیشنهادات را به طرز چشم‌گیری افزایش دهد. یکی از راه‌های یافتن همسایه‌ها، استفاده از معیارهای اندازه‌گیری شباهت می‌باشد. اندازه‌گیری شباهت، از امتیازات آیتم‌های مشترک، برای محاسبه شباهت بین کاربر فعال و سایر کاربران استفاده می‌کند. ضریب همبستگی پیرسون که در تحقیقات اخیر به طور گسترده‌ای استفاده شده است، دارای معایبی همچون عدم لحاظ کردن تعداد آیتم های مشترک، عدم لحاظ کردن فاصله امتیازات، عدم لحاظ کردن ارزش امتیازات می‌باشد. در این تحقیق و در روش پیشنهادی، این معایب بر طرف شده است و یک ضریب جدید با استفاده از ضریب همبستگی پیرسون معرفی شده است. ضریب پیشنهادی شبیه‌ترین افراد به یکدیگر را پیدا کرده و خطای پیشنهادات را به طرز چشم‌گیری کاهش می‌دهد. برای بررسی عملکرد و ارزیابی نتایج روش پیشنهادی، از دو مجموعه داده **MovieLens100k** و **Jester** استفاده کرده‌ایم. در تمامی نتایج شبیه سازی، روش پیشنهادی دارای کمترین خطا نسبت به سایر روش‌های موجود می‌باشد.

**واژگان کلیدی:** سیستم‌های پیشنهادگر- پالایش مشارکتی- معیارهای شباهت- ضریب همبستگی پیرسون

مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۵ ، بهار ۱۳۹۶، ص ۱۹- ۸

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

**مقدمه**

امروزه اینترنت بخش مهمی از زندگی افراد را شامل می‌شود. همچنین توسعه دهندگان وب نیز انواع مختلفی از داده‌ها را به صورت آنلاین در اختیار مردم قرار می‌دهند و هر روزه نیز بر مقدار این اطلاعات افزوده می‌شود [1] . بنابراین افراد تقریبا خود را در انبوهی از اطلاعات احاطه می‌بینند. در این انبوه اطلاعات، موتورهای جستجو تا حدی به یافتن اطلاعات مورد نیاز به ما کمک می‌کنند، با این حال، اگر سیستمی که بتواند بدون نیاز و درخواست از آن بتواند به صورت خودمختار پیشنهاد و توصیه‌های به ما داشته باشد بسیار مفید خواهد بود [2,6].

فرض کنید به دنبال خرید یک دستگاه لب تاپ هستید، احتمالاً برای انجام این کار با گزینه‌های مختلفی روبه‌رو می‌شوید و شاید دچار سردرگمی شوید و حتی قادر به تصمیم‌گیری نیز نباشید. سابق بر این، برای غلبه بر این نوع مشکلات احتمالاً از نظرات دوستان، همکلاسی‌ها و یا همکاران خود کمک می‌گرفتید، ولی امروزه هیچ انسانی نمی‌تواند ادعا کند که با توجه به تمامی اطلاعات موجود به شما پیشنهادات مختلف را ارائه می‌دهد. امروزه استفاده از سیستم‌های پیشنهادگر به‌عنوان یک ضرورت در آمده است و بسیاری از سایت‌های اینترنتی برای خدمت‌رسانی شایسته به مشتریان خود و فروش بیشتر محصولات خود، از سیستم پیشنهادگر استفاده می‌کنند [6,7]. سیستم‌های پیشنهادگر می‌توانند براساس سلایق افراد شخصی‌سازی شده و بهترین آیتم‌ها را به افراد پیشنهاد دهند. سیستم‌های پیشنهادگر در بسیاری از زمینه‌ها همچون توریسم [8]، فیلم [3,9]، موزیک [10]، اخبار [2,6] به کار رفته‌اند. از جمله بارزترین آن‌ها می‌توان به MovieLens, Amazon، FaceBook اشاره نمود. سامانه پیشنهادگر، با تحلیل رفتار کاربر خود، اقدام به پیشنهاد مناسب‌ترین اقلام (داده، اطلاعات، کالا و...) می‌نماید. این سیستم رویکردی است که برای مواجهه با مشکلات ناشی از حجم فراوان و رو به رشد اطلاعات ارائه شده‌است و به کاربر خود کمک می‌کند تا در میان حجم عظیم اطلاعات سریع‌تر به هدف خود نزدیک شوند. بسیاری از سیستم‌های پیشنهادگر موفق در وب از روش پالایش مشارکتی استفاده می‌کنند[11]. هدف اصلی روش پالایش مشارکتی پیشنهاد آیتم‌ها به کاربر فعال براساس کاربران همسایه می‌باشد. روش‌های پالایش مشارکتی به دو دسته شامل مدل محور و حافظه محور تقسیم‌بندی می‌شوند. مدل محور شامل روش‌های می‌باشد که در فاز آفلاین به یادگیری یک مدل می‌پردازد و به صورت آنلاین از این مدل برای عمل توصیه استفاده می‌شود. از طرف دیگر، دسته حافظه محور از روش‌های اکتشافی برای توصیه استفاده می‌کنند. در این روش‌ها، توصیه براساس امتیازات دردسترس به آیتم‌ها صورت می‌گیرد. در روش‌های حافظه محور یک مجموعه m کاربر $U=\{u_1,u_2,...,u_m\}$ و n مجموعه از آیتم‌ها $I=\{i_1,i_2,...,i_n\}$ در دسترس می‌باشد. هرکابر علاقه خود را به آیتم‌های دیده شده به وسیله مقدار در مقایس تعریف شده نسبت می‌دهد. این علاقه‌ها توسط ماتریس m*n که ماتریس امتیازات نامیده می‌شود، نشان داده می‌شود [11]. ایده اصلی روش پالایش مشارکتی به این صورت است: اگر دو کاربر امتیازدهی‌های یکسان بر روی آیتم‌های مشترک داشته باشند، انگاه آن‌ها علاقه-های یکسانی دارند. بنابراین، در این روش، پیشنهادات به کابر فعال براساس کاربران همسایگان انجام می‌شود [11]. اندازه گیری شباهت و الگوریتم های خوشه بندی دو راه برای پیدا کردن همسایه‌ها می‌باشند. یکی از کلیدی‌ترین بخش از سیستم‌های پیشنهادگر، بخش پیدا کردن همسایه‌ها می‌باشد که اگر به درستی انتخاب شود می‌تواند صحت پیشنهادات را به طرز چشم‌گیری افزایش دهد. هدف از الگوریتم-های خوشه‌بندی گروه‌بندی کاربران در تعدادی خوشه براساس الگوهای مشترک‌شان می‌باشد. کاربرانی که در یک خوشه با کاربر فعال قرار دارند به عنوان همسایگان انتخاب می‌شوند. K-Means به‌صورت گسترده به‌عنوان الگوریتم خوشه‌بندی در چندین روش پالایش مشارکتی استفاده شده است[11].

اندازه‌گیری شباهت، از امتیازات آیتم‌های مشترک، برای محاسبه شباهت بین کاربر فعال و سایر کاربرین استفاده می‌کند. چندین اندازه‌گیری شباهت، برای محاسبه شباهت در کارها گزارش شده است. ضریب همبستگی پیرسون و روش شباهت کسینوسی به-صورت گسترده برای اندازه‌گیری شباهت استفاده شده است[12,13]. اخیرا نیز معیارهای شباهت دیگری نیز معرفی شده است. به عنوان مثال در پژوهش [14] از مزایای سیستم‌های فازی استفاده شده است و معیار شباهتی با عنوان معیار شباهت وزن‌دهی فازی، معرفی کرده است. معیار معرفی شده در کار [14] فقط ارزش امتیازات را در ضریب همبستگی پیرسون لحاظ کرده است و تعداد امتیازات مشترک و فاصله امتیازات را در میزان شباهت لحاظ نکرده است. در کار [15] از ضریب جاکارد استفاده کرده است و معیار شباهت جدیدی با عنوان ضریب توانی معرفی کرده است. به دلیل اینکه ضریب جاکارد بر روی مجموعه‌های دودویی به کار می‌رود استفاده از این ضریب در این حوزه کارایی لازم را ندارد.

به عنوان مثال نویسنده [15] امتیازات را به دو مجموعه امتیازات خوب و امتیازات بد افراز کرده است. یک تقسیم‌بندی امتیازات به شکل زیر می‌باشد:

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

$$G = \{g_1, ..., g_c\}$$

$$B = \{b_1, ..., b_d\}$$

در این افراز، $g_1, ..., g_c$ در یک مجموعه قرار دارند و هیچ تفاوتی بین امتیازات در این گروه قرار ندارد. این رویکرد، میزان خطای پیش بینی را افزایش خواهد داد.

## ۱- انواع معیارهای شباهت

### ۱-۱ ضریب همبستگی پیرسون

این ضریب میزان همبستگی بین دو متغیر فاصله‌ای یا نسبی را محاسبه کرده مقدار آن بین ۱+ و ۱- می‌باشد، اگر مقدار بدست آمده مثبت باشد به معنی این است که تغییرات دو متغیر به طور هم جهت اتفاق می‌افتد یعنی با افزایش در هر متغیر، متغیر دیگر نیز افزایش می‌یابد و برعکس، اگر مقدار بدست آمده منفی شد، یعنی اینکه دو متغیر در جهت عکس هم عمل می‌کنند، یعنی با افزایش مقدار یک متغیر مقادیر متغیر دیگر کاهش می‌یابد و برعکس. اگر مقدار بدست آمده صفر شد نشان می‌دهد که هیچ رابطه‌ای بین دو متغیر وجود ندارد و اگر ۱+ شد همبستگی مثبت کامل و اگر ۱- شد همبستگی کامل و منفی است. مقدار این ضریب توسط رابطه (۱) محاسبه می-شود.

$$PCC_{u_i,u_j} = \frac{\left( \sum_{\alpha \in O_{u_i u_j}} (r_{u_j \alpha} - \bar{r}_{u_j})(r_{u_i \alpha} - \bar{r}_{u_i}) \right)}{\left( \sqrt{\sum_{\alpha \in O_{u_i u_j}} (r_{u_j \alpha} - \bar{r}_{u_j})^2} \sqrt{\sum_{\alpha \in O_{u_i u_j}} (r_{u_i \alpha} - \bar{r}_{u_i})^2} \right)} \quad (۱)$$

که $O_{u_i u_j}$ مجموعه آیتم‌های مشترک کاربر $u_i$ و $u_j$ می‌باشد و $r_{u_j}$ میانگین امتیاز کاربر $u_j$ می‌باشد. $r_{u_i}$ میانگین امتیاز کاربر $u_i$ می‌باشد. ماتریس رتبه‌دهی جدول(۱) را در نظر بگیرید.

**جدول۱- ماتریس رتبه دهی**

| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 |
|---|---|---|---|---|---|
| User 1 | 1 | - | 2 | 3 | 4 |
| User 2 | 1 | - | 2 | 3 | 5 |
| User 3 | 1 | 1 | 3 | - | - |
| User 4 | 2 | - | 3 | 4 | 5 |
| User 5 | 2 | 2 | - | 2 | - |

با استفاده از ضریب همبستگی پیرسون مقدار شباهت کاربران به صورت جدول(۲) محاسبه شده است.

**جدول۲- شباهت میان کاربران**

| | **User1** | **User2** | **User3** | **User4** | **User5** |
|---|---|---|---|---|---|
| **User1** | - | 0.9827 | 1.0 | 1.0 | Nan |
| **User2** | | - | 1.0 | 0.9827 | Nan |
| **User3** | | | - | 1.0 | Nan |
| **User4** | | | | - | Nan |
| **User5** | | | | | - |

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

با توجه به جدول(۲) مقدار شباهت کاربران Sim(U1,U2)<Sim(U1,U3) بدست آمده است، در حالیکه کاربران U1,U2 شباهت بیشتری نسبت به کاربران U1,U3 دارند. پس می‌توان نتیجه گرفت یکی از معایب ضریب همبستگی پیرسون عدم لحاظ کردن تعداد آیتم‌های مشترک در میان دو کاربر می‌باشد.

همچنین با دقت به جدول(۲) می‌توان مشاهده کرد Sim(U1,U2)<Sim(U1,U4) بدست آمده است. کاربران U1,U4 در هیچ آیتمی با یکدیگر هم نظر نبوده‌اند اما مقدار شباهت آن‌ها 1.0 شده است و از U1,U2 نیز بیشتر شده‌اند. علت این تناقض آشکار این است که ضریب همبستگی پیرسون فقط به رشد اعداد توجه نموده است و فاصله امتیازات در محاسبه شباهت میان کاربران در نظر گرفته نشده است. به دلیل اینکه میانگین رتبه کاربر شماره پنجم 2 می‌باشد مقدار شباهت بدست آمده برای کاربر U5 با سایر کاربرین تعریف نشده بدست آمده است. از معایب دیگر روش پیرسون نیز می‌توان به عدم لحاظ کردن ارزش و وزن برای آیتم‌ها اشاره کرد.

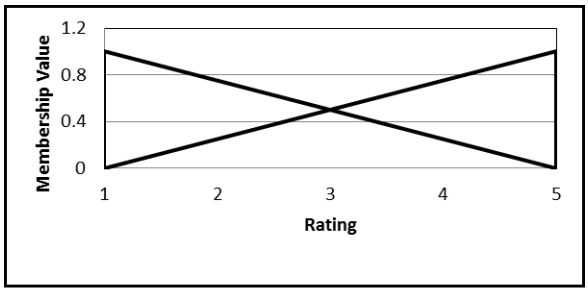
### ۲-۱ شباهت کسینوسی

معیار شباهت کسینوسی یک معیار اندازه‌گیری شباهت میان دو بردار می‌باشد که کسینوس زاویه بین آن دو بردار را اندازه‌گیری می‌کند[14,15]. کسینوس صفر درجه مقدار ۱+ می‌شود و مقدار کمتر از ۱+ برای سایر زوایا می‌باشد. بنابراین شباهت کسینوسی براساس جهت دو بردار میزان شباهت را بدست می‌آورد و به دامنه دو بردار توجه‌ای ندارد. دو بردار با جهات یکسان دارای شباهت کسینوسی ۱+ می‌باشند. دو بردار با زاویه ۹۰ درجه دارای شباهت ۰ می‌باشند و دو بردار با جهت‌های مخالف دارای شباهت ۱- می‌باشند. لازم به ذکر است شباهت کسینوسی مستقل از دامنه دو بردار می‌باشد. میزان شباهت کسینوسی توسط رابطه (۲) محاسبه می گردد.

$$Cosine\ Similarity_{u_i,u_j} = \frac{\left(\sum_{\alpha \in O_{u_i u_j}} (r_{u_j \alpha} \cdot r_{u_i \alpha})\right)}{\left(\sqrt{\sum_{\alpha \in O_{u_i u_j}} (r_{u_j \alpha})^2} \sqrt{\sum_{\alpha \in O_{u_i u_j}} (r_{u_i \alpha})^2}\right)} \quad (۲)$$

### ۳-۱ شباهت وزن دهی با استفاده از سیستم‌های فازی

یکی از معایبی که استفاده از الگوریتم ژنتیک به عنوان معیار شباهت با چالش روبرو می‌نماید بحث زمان می‌باشد[14]، در صورتی که در سیستم‌های پیشنهادگر اغلب، حداقل زمان پاسخ‌گویی مد نظر می‌باشد. در کار[14] برای یافتن وزن آیتم مشترک از مزایای سیستم‌های فازی استفاده می‌نماید. در ابتدا امتیازات از طریق توابع عضویت به مقادیر فازی نگاشت داده می‌شوند. توابع عضویت برای نگاشت به مقادیر فازی به‌صورت شکل(۱) تعریف شده است.




**شکل ۱- مجموعه فازی پیشنهادی برای مقادیر امتیازات در کار[24]**

به‌عنوان مثال اگر امتیاز کاربر اول به آیتم اول ۲ باشد، آنگاه مقدار تابع عضویت مقادیر $<0.25,0.75>$ خواهد شد. سپس براساس مقادیر تابع عضویت وزن امتیاز آیتم مشترک $\alpha$ به صورت رابطه(۳) محاسبه می‌شود:

$$w_\alpha = \sqrt{2} - dis(\tilde{r}_{i\alpha} - \tilde{r}_{j\alpha}) \quad (۳)$$

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

$dis(\tilde{r}_{i\alpha}-\tilde{r}_{j\alpha})$ هرگونه متریک فاصله می‌تواند باشد. در کار[14] از فاصله اقلیدسی به‌عنوان اندازه‌گیری فاصله استفاده شده است و مقدار آن توسط رابطه (۴) محاسبه می‌گردد. $\tilde{r}_{i\alpha}$ و $\tilde{r}_{j\alpha}$ ۲ بردار عضویت با سایز $l$ می‌باشد.

$$dis(\tilde{r}_{i\alpha}-\tilde{r}_{j\alpha})=\sqrt{\sum_{n=1}^{l}(\tilde{r}_{i\alpha}^{n}-\tilde{r}_{j\alpha}^{n})^2} \quad (۴)$$

$\tilde{r}_{i\alpha}^{n}$ و $\tilde{r}_{j\alpha}^{n}$ مقدار عضویت مجموعه فازی nم می‌باشد. مقدار $dis(\tilde{r}_{i\alpha}-\tilde{r}_{j\alpha})$ از $\sqrt{2}$ در فرمول (۳) کم شده است زیرا بیشترین فاصله است که از فرمول (۴) به‌دست می‌آید. ( به عنوان مثال اگر $\tilde{r}_{i\alpha}=<0,1>$ و $\tilde{r}_{j\alpha}=<1,0>$ باشد آنگاه فاصله آنها $\sqrt{2}$ خواهد شد و درنتیجه وزن نهایی 0 می‌شود) و از فرمول (۵) برای میزان شباهت میان کاربران استفاده شده است.

$$Corr(u_i,u_j)=\left(\sum_{\alpha\in O_{u_iu_j}} w_\alpha(r_{u_j\alpha}-\bar{r}_{u_j})(r_{u_i\alpha}-\bar{r}_{u_i})\right)\Big/\left(\sqrt{\sum_{\alpha\in O_{u_iu_j}}(r_{u_j\alpha}-\bar{r}_{u_j})^2}\sqrt{\sum_{\alpha\in O_{u_iu_j}}(r_{u_i\alpha}-\bar{r}_{u_i})^2}\right) \quad (۵)$$

**۴-۱ شباهت ضریب توانی**

در کار[15] نویسنده از ضریب جاکارد استفاده کرده و معیار شباهتی با عنوان ضریب توانی معرفی کرده است. ضریب جاکارد بر روی مجموعه‌های دودویی اعمال می‌شود و از تقسیم تعداد اشتراک دو مجموعه بر تعداد اجتماع دو مجموعه بدست می‌آید.

$$J_p(E,F)=\frac{|E\cap F|}{|E\cup F|}=\frac{p}{p+q+r} \quad (۶)$$

P تعداد تطابق هر دو مثبت در مجموعه E,F می‌باشد. q تعداد تطابقی که در مجموعه E مثبت و در مجموعه F منفی می‌باشند.r تعداد تطابقی که در مجموعه E منفی و در مجموعه F مثبت می‌باشند. نحوه محاسبه p,q,r,n در جدول نشان داده شده است. همان طور که عنوان شد برای استفاده از ضریب جاکارد نیاز است که امتیازات را به صورت دودویی تقسیم بندی نمود. به عبارت دیگر نویسنده دو مجموعه خوب و بد را به صورت زیر تعریف نموده است.

$$G=\{g_1,...,g_c\}, B=\{b_1,...,b_d\}$$

d,c به ترتیب تعداد امتیازات خوب و بد می‌باشد. یکی از معایب ضریب جاکارد لحاظ نکردن تطابق منفی می‌باشد. برای حل این مشکل نویسنده [15] تطابق‌های منفی را نیز به صورت رابطه (۷) در نظر گرفته است.

$$J_n(E,F)=\frac{n}{n+q+r} \quad (۷)$$

در مرحله بعد با میانگین‌گیری از روابط (۶) و (۷) شباهت به صورت رابطه (۸) به دست آورده است.

$$sim(E,F)=\frac{J_p(E,F)+J_n(E,F)}{2} \quad (۸)$$

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

نویسنده برای افزایش تاثیر تطابق‌های مثبت از رابطه (۹) استفاده کرده است.

$$P_p^{\alpha}(E,F)=\frac{p^{\alpha}}{p^{\alpha}+q+r} \quad (۹)$$

در رابطه بالا درجه $\alpha$ اهمیت تطابق های مثبت را نشان می‌دهد. اگر باشد همان ضریب جاکارد را خواهیم داشت. به طور مشابه برای افزایش تاثیر تطابق های منفی نیز از رابطه (۱۰) استفاده می‌شود.

$$P_n^{\alpha}(E,F)=\frac{n^{\alpha}}{n^{\alpha}+q+r} \quad (۱۰)$$

در نهایت با میانگین‌گیری از روابط (۹)و (۱۰) میزان شباهت نهایی توسط رابطه (۱۱) بدست می‌آید.

$$sim(E,F)=\frac{P_p^{\alpha}(E,F)+P_n^{\alpha}(E,F)}{2} \quad (۱۱)$$

## ۲- روش پیشنهادی

انتخاب صحیح یک تابع شباهت، برای تعیین شباهت میان کاربران یک فاکتور مهم و حیاتی در الگوریتم پالایش مشارکتی است، زیرا صحت پیشنهادات را به شدت تحت تاثیر قرار می‌دهد. مطالعات در این زمینه ثابت کردند که ضریب همبستگی پیرسون بهتر از سایر معیارهای شباهت عمل می‌کند [1] .این ضریب، رابطه خطی بین دو متغیر مشخص را می‌کند. و مقدار آن از ۱- تا ۱+ متغیر است. مقدار ۱+ نشان دهنده ارتباط کامل دو متغیر و مقدار ۱- نمایش‌دهنده عدم ارتباط دو متغیر است. به‌عبارت دیگر ۱+ نمایش می‌دهد که دو کاربر کاملا علایق مرتبط با هم دارند در صورتی که عدد ۱- نمایش دهنده تضاد علایق دو کاربر است ضریب همبستگی پیرسون به صورت گسترده به عنوان معیار شباهت در سیستم‌های پیشنهادگر استفاده شده است که دارای معایبی می باشد. برای لحاظ کردن تعداد آیتم‌های مشترک از رابطه (۱۲) به عنوان ضریب استفاده خواهیم کرد.

$$\frac{\left|O_{u_i u_j}\right|}{|I|} \quad (۱۲)$$

رابطه (۱۲) نسبت تعداد آیتم‌های مشترک به تعداد کل آیتم ها می‌باشد که $\left|O_{u_i u_j}\right|$ تعداد آیتم‌های مشترک می‌باشد و $|I|$ تعداد کل آیتم‌ها می‌باشد. برای برطرف نمودن مشکل، عدم لحاظ کردن فاصله امتیازات از ضریب (۱۳) استفاده خواهیم کرد.

$$\frac{1}{1+\sqrt{\sum_{i=1}^{m}(r_{a,i}-r_{b,i})^2}} \quad (۱۳)$$

در صورت یکسان بودن امتیازات ۲کاربر مقدار این ضریب ۱ خواهد بود و در صورت یکسان نبودن امتیازات مقدار کمتر از ۱ خواهد داشت. به منظور بر طرف کردن مشکل ارزش آیتم ها از کار [14] استفاده خواهیم کرد. در کار [14] از مزایای سیستم‌های فازی استفاده کرده است و ارزش آیتم ها را به خوبی بدست آورده است. در نهایت معیار پیشنهادی این پژوهش به صورت رابطه(۱۴) شباهت میان کاربران را بدست می‌آورد.

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

$$Similarity\,(u_i,u_j)=\frac{\left(\sum_{\alpha\in O_{u_iu_j}} W_\alpha(r_{u_j\alpha}-\bar{r}_{u_j})(r_{u_i\alpha}-\bar{r}_{u_i})\right)}{\left(\sqrt{\sum_{\alpha\in O_{u_iu_j}}(r_{u_j\alpha}-\bar{r}_{u_j})^2}\sqrt{\sum_{\alpha\in O_{u_iu_j}}(r_{u_i\alpha}-\bar{r}_{u_i})^2}\right)}*\left(\frac{1}{1+\sqrt{\sum_{i=1}^{m}(r_{a,i}-r_{b,i})^2}}\right)*\frac{|O_{u_iu_j}|}{|I|} \quad (۱۴)$$

## ۳- ارزیابی نتایج

### ۳-۱ مجموعه داده‌های آزمایش

برای ارزیابی عملکرد روش پیشنهادی از دو مجموعه داده MovieLens100K و Jester استفاده کرده‌ایم. در جدول زیر مشخصات این مجموعه داده‌ها را مشاهده می‌کنید.

**جدول ۳- مجموعه داده‌های استفاده شده برای بررسی عملکرد روش پیشنهادی**

| مجموعه داده | تعداد کاربر | تعداد آیتم | نوع آیتم | تعداد امتیازات | مقیاس امتیازات | نوع امتیازات |
|---|---|---|---|---|---|---|
| MovieLens100K | ۹۴۳ | ۱۶۸۲ | فیلم | ۱۰۰۰۰۰ | ۱+ تا ۵+ | صحیح |
| Jester | ۲۰۰۰ | ۱۰۰ | جوک | ۱۴۴۶۵۹ | ۱۰- تا ۱۰+ | حقیقی |

برای بررسی عملکرد روش پیشنهادی ، داده‌ها به دو مجموعه داده‌های آموزشی و داده‌های آزمایشی تقسیم شده‌اند که مجموعه آموزشی شامل ۸۰٪ داده‌ها و مجموعه آزمایشی شامل ۲۰٪ داده ها می‌باشد.

### ۳-۲ معیارهای ارزیابی

#### ۱) میانگین مطلق خطا

رایج‌ترین معیار برای مقایسه، میانگین مطلق خطا است که برای ارزیابی توانایی یک سیستم جهت پیش بینی علاقه یک کاربر به آیتمی خاص به کار می‌رود. این معیار، میانگین قدر مطلق تفاضل‌های مقدار واقعی و مقدار پیش‌بینی شده را به صورت زیر بدست می‌آورد[14,15,16].

$$MAE=\frac{\sum_{i=1}^{n}|r_{a,i}-r_{p,i}|}{n} \quad (۱۵)$$

که در آن $r_{a,i}$ امتیاز واقعی کاربر a به آیتم i می‌باشد. $r_{p,i}$ امتیاز پیش‌بینی شده کاربر a به آیتم i می‌باشد و N تعداد پیش‌بینی‌ها می‌باشد. معیار MAE به‌طور گسترده در ارزیابی سیستم‌های پیشنهادگر استفاده شده است.

#### ۲) مجذور میانگین مربعات خطا

مجذور میانگین مربعات خطا نیز یکی از معیارهای ارزیابی در سیستم‌های پیشنهادگر می‌باشد که به صورت رابطه (۱۶) تعریف می‌گردد[16].

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}|r_{a,i}-r_{p,i}|}{n}} \quad (۱۶)$$

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

$r_{a,i}$ امتیاز واقعی کاربر a به آیتم i می‌باشد. $r_{p,i}$ امتیاز پیش‌بینی شده کاربر a به آیتم i می‌باشد و N تعداد پیش‌بینی-ها می‌باشد.

### ۳) درصد صحیح موارد پیش بینی شده

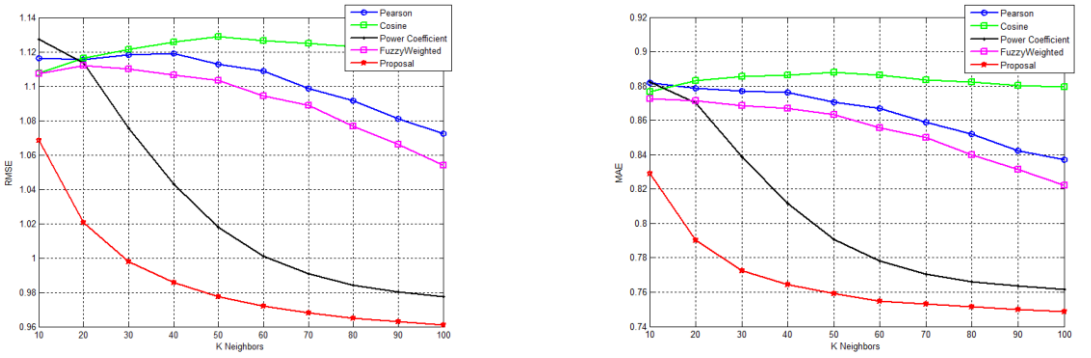
معیار رایج دیگری که برای ارزیابی سیستم‌های پیشنهادگر استفاده می‌شود PCP می‌باشد[14,15] و به‌صورت رابطه(۱۷) تعریف می‌گردد.

$$CorrectSet(u_a) = \{i_k \mid i_k \in I_{Test}, r_{a,i} = r_{p,i}\} \quad (۱۷)$$

$$PCP = \frac{\sum_{i=1}^{n} |CorrectSet(u_i)|}{\sum_{i=1}^{n} |I_i^{Test}|} \times 100\%$$

## ۴- نتایج شبیه‌سازی

در شکل(۲) مقادیر میانگین مطلق خطا و مجذور میانگین مربعات خطا برای مجموعه داده MovieLens100K نشان داده شده است.همان طور که از شکل(۲) نمایان است روش پیشنهادی به ازای تمام مقادیر همسایگان دارای کمترین مقدار خطا نسبت به سایر روش‌ها دارا می‌باشد.



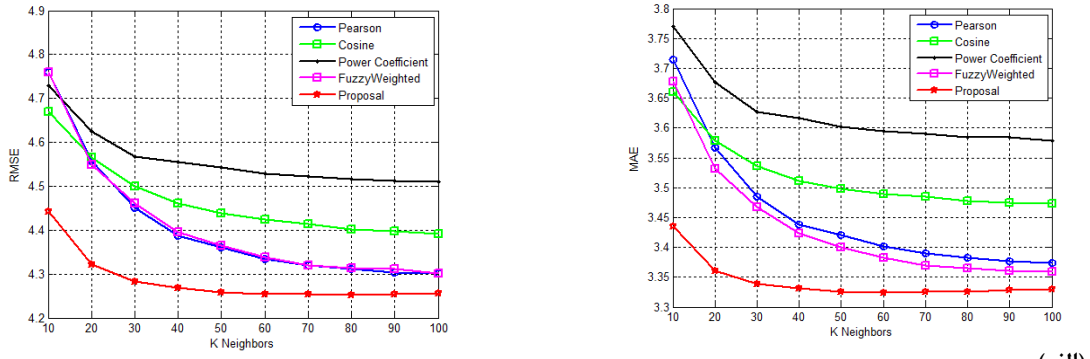

(ب) (الف)

**شکل۲- (الف) میانگین مطلق خطا و (ب) مجذور میانگین مربعات خطا برای مجموعه داده MoveiLens100k**

همان‌طور که از شکل (۲) مشاهده می‌شود، لحاظ کردن تعداد، ارزش و فاصله امتیازات در ضریب همبستگی پیرسون، میزان شباهت افراد را به صورت بهینه محاسبه کرده است و خطای پیشنهادات در مجموعه داده MoveiLens100k به طرز چشم‌گیری کاهش داده است.

در شکل(۳) نیز مقادیر میانگین مطلق خطا و مجذور میانگین مربعات خطا برای مجموعه داده Jester نشان داده شده است.همان طور که از شکل(۳) نمایان است روش پیشنهادی نیز در مجموعه داده Jester دارای کمترین مقدار خطا می‌باشد.

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




**شکل۳- (الف)میانگین مطلق خطا و (ب) مجذور میانگین مربعات خطا برای مجموعه داده Jester**

با توجه به شکل(۳) روش پیشنهادی نیز در مجموعه داده Jester خطای پیشنهادات را کاهش داده است. در جدول(۴) مقادیر PCP روش پیشنهادی و سایر روش‌ها در مجموعه داده MovieLens100k نشان داده شده است. بهترین مقادیر نتایج در بین روش‌ها با فونت ضخیم نشان داده شده است. با توجه به جدول(۴) آشکار است که روش پیشنهادی در سایز همسایه‌های کم، درصد موارد پیش بینی صحیح بالاتری نسبت به سایر کارها داشته است و در تمامی همسایه‌ها نیز درصد صحیح موارد پیش بینی بالاتر از سایر روش‌ها می-باشد.

**جدول ۴- مقایسه مقادیر PCP در مجموعه داده MovieLens100k**

| Neighbor Size | Pearson | Cosine | Fuzzy Weighted | Power Coefficient | Proposal |
|---|---|---|---|---|---|
| 10 | 34.805 | 34.605 | 35.01 | 35.46 | 38.13 |
| 20 | 35.075 | 34.55 | 35.635 | 35.615 | 40.055 |
| 30 | 35.355 | 34.625 | 35.825 | 37.455 | 40.545 |
| 40 | 35.39 | 34.815 | 35.785 | 38.51 | 41.135 |
| 50 | 35.575 | 34.795 | 36.185 | 39.875 | 41.315 |
| 60 | 35.92 | 34.93 | 36.615 | 40.215 | 41.525 |
| 70 | 36.23 | 34.98 | 36.97 | 40.775 | 41.43 |
| 80 | 36.815 | 35.17 | 37.325 | 40.945 | 41.495 |
| 90 | 37.32 | 35.27 | 37.735 | 41.025 | 41.615 |
| 100 | 37.505 | 35.325 | 38.15 | 41.075 | 41.58 |

یکی از کلیدی‌ترین اجزا در سیستم‌های پیشنهادگر، بخش پیدا کردن همسایه‌های کاربر فعال می‌باشد که اگر به‌درستی انتخاب شود، می‌تواند صحت پیشنهادات را به طرز چشم‌گیری افزایش دهد. دلیل برتری روش پیشنهادی را می‌توان به یافتن صحیح‌تر کاربران همسایه کاربر فعال اشاره کرد. در جدول(۵) نیز مقادیر PCP برای مجموعه داده Jester نشان داده شده است. بهترین مقادیر PCP در بین روش ها به صورت فونت درشت نشان داده شده است. با توجه به جدول مشاهده می‌گردد درصد صحیح موارد پیش بینی روش پیشنهادی در تمامی همسایه‌ها دارای بالاترین مقدار نسبت به سایر روش‌ها می‌باشد.

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

**جدول ۵- مقایسه مقادیر PCP در مجموعه داده Jester**

| Neighbor Size | Pearson | Cosine | Fuzzy Weighted | Power Coefficient | Proposal |
|---|---|---|---|---|---|
| 10 | 9.086 | 9.932 | 9.639 | 8.858 | 10.130 |
| 20 | 9.370 | 9.729 | 9.740 | 8.706 | 10.286 |
| 30 | 9.653 | 9.408 | 9.850 | 8.675 | 10.210 |
| 40 | 9.833 | 9.584 | 9.905 | 8.748 | 10.327 |
| 50 | 9.798 | 9.539 | 10.061 | 8.779 | 10.358 |
| 60 | 9.712 | 9.435 | 9.992 | 8.806 | 10.279 |
| 70 | 9.937 | 9.314 | 10.120 | 8.996 | 10.258 |
| 80 | 10.040 | 9.332 | 10.096 | 9.031 | 10.351 |
| 90 | 10.096 | 9.418 | 10.201 | 9.076 | 10.241 |
| 100 | 10.154 | 9.460 | 10.230 | 9.072 | 10.255 |

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

## ۵- نتیجه گیری

ایده اصلی روش پالایش مشارکتی به این صورت است که اگر دو کاربر امتیازدهی‌های یکسان بر روی آیتم‌های مشترک داشته باشند، انگاه آن‌ها علاقه‌های یکسانی دارند. بنابراین، در این روش، پیشنهادات به کابر فعال براساس کاربران همسایگان انجام می‌شود. یکی از کلیدی‌ترین بخش از سیستم‌های پیشنهادگر، بخش پیدا کردن همسایه‌ها می‌باشد که اگر به‌درستی انتخاب شود می‌تواند صحت پیشنهادات را به طرز چشم‌گیری افزایش دهد. یکی از راه‌های یافتن همسایه‌ها، استفاده از معیارهای اندازه‌گیری شباهت می‌باشد. اندازه-گیری شباهت، از امتیازات آیتم‌های مشترک، برای محاسبه شباهت بین کاربر فعال و سایر کاربرین استفاده می‌کند. ضریب همبستگی پیرسون به‌صورت گسترده برای اندازه‌گیری شباهت استفاده شده است. ضریب همبستگی پیرسون دارای معایبی همچون عدم لحاظ کردن تعداد آیتم‌های مشترک، عدم لحاظ کردن فاصله امتیازات، عدم لحاظ کردن ارزش امتیازات می‌باشد. در روش پیشنهادی، این معایب بر طرف شده است و یک معیار جدید با استفاده از ضریب همبستگی پیرسون معرفی شده است. برای بررسی عملکرد روش پیشنهادی ما از دو مجموعه داده MovieLens100k و Jester استفاده کرده‌ایم. در تمامی نتایج شبیه سازی روش پیشنهادی دارای کمترین خطا می-باشد. لازم به ذکر است از روش پیشنهادی نیز می‌توان در سایر کاربردها بهره جست.

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

## منابع و مراجع

[1] Phelps, J. E., Lewis, R., Mobilio, L., Perry, D., & Raman, N. (2004). Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email. Journal of advertising research, 44(04), 333-348.
[2] Wen, H., Fang, L., & Guan, L. (2012). A hybrid approach for personalized recommendation of news on the Web. Expert Systems with Applications, 39(5), 5806-5814.
[3] Choi, S. M., Ko, S. K., & Han, Y. S. (2012). A movie recommendation algorithm based on genre correlations. Expert Systems with Applications, 39(9), 8079-8085.
[4] Tsai, C. F., & Hung, C. (2012). Cluster ensembles in collaborative filtering recommendation. Applied Soft Computing, 12(4), 1417-1425.
[5] Luo, X., Xia, Y., & Zhu, Q. (2012). Incremental collaborative filtering recommender based on regularized matrix factorization. Knowledge-Based Systems, 27, 271-280.
[6] Cleger-Tamayo, S., Fernández-Luna, J. M., & Huete, J. F. (2012). Top-N news recommendations in digital newspapers. Knowledge-Based Systems, 27, 180-189.
[7] Guan, Y., Zhao, D., Zeng, A., & Shang, M. S. (2013). Preference of online users and personalized recommendations. Physica A: Statistical Mechanics and its Applications, 392(16), 3417-3423.
[8] Garcia, I., Sebastia, L., & Onaindia, E. (2011). On the design of individual and group recommender systems for tourism. Expert systems with applications, 38(6), 7683-7692.
[9] Carrer-Neto, W., Hernández-Alcaraz, M. L., Valencia-García, R., & García-Sánchez, F. (2012). Social knowledge-based recommender system. Application to the movies domain. Expert Systems with applications, 39(12), 10990-11000.
[10] Li, Q., Myaeng, S. H., & Kim, B. M. (2007). A probabilistic music recommender considering user opinions and audio features. Information processing & management, 43(2), 473-487.
[11] Ramezani, M., Moradi, P., & Akhlaghian, F. (2014). A pattern mining approach to enhance the accuracy of collaborative filtering in sparse data domains. *Physica A: Statistical Mechanics and its Applications*, *408*, 72-84.
[12] Liu, H., Hu, Z., Mian, A., Tian, H., & Zhu, X. (2014). A new user similarity model to improve the accuracy of collaborative filtering. *Knowledge-Based Systems*, *56*, 156-166.
[13] Nilashi, M., & Ibrahim, O. B. (2014). A model for detecting customer level intentions to purchase in B2C websites using TOPSIS and fuzzy logic rule-based system. *Arabian Journal for Science and Engineering*, *39*(3), 1907-1922.
[14] Al-Shamri, M. Y. H., & Al-Ashwal, N. H. (2014). Fuzzy-weighted similarity measures for memory-based collaborative recommender systems. *Journal of Intelligent Learning Systems and Applications*, *2014*.
[15] Al-Shamri, M. Y. H. (2014). Power coefficient as a similarity measure for memory-based collaborative recommender systems. *Expert Systems with Applications*, *41*(13), 5680-5688.
[16] Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, *41*(4), 2065-2073.