7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

Enhancing Tag-Based Recommender System Using Ontology

Elahe Nosouhi[1], Sasan Hoseinalizadeh[2]

[1]Master of Information Technology Engineering, Qazvin Islamic Azad University, Qazvin, Iran
enosouhi@gmail.com

[2]Assistant Professor, Communication and Information Technology Research Institute, Tehran, Iran
s.alizadeh@itrc.ac.ir

Abstract

Due to the information overload on the World Wide Web, the user suffers from difficulty in selecting items. Social cataloging services allow users to use products or services and share their opinions and experiences, which are effective not only for themselves, but also for other users. Considering user behavior and product features as the two determining factors, recommender systems have significantly influenced the item selection process. In this paper, according to the emotions reflected in the user tags, a tag-based recommendation method is proposed. The method works in the following way: information related to these emotions, along with other information received from the user as well as the content information of the items results in obtaining the degree of similarity between them. This process ultimately helps to improve the performance of the recommender systems. Testing the abovementioned process on a real database, namely Movielense, showed that the proposed method performed better than previous ones and has reduced errors and increased accuracy in predicting ratings.

Keywords: Recommender system, Tag, Ontology, Rating prediction.

7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

بهبود سیستم‌های توصیه‌گر مبتنی بر برچسب با استفاده از آنتولوژی

*الهه نصوحی[1]، ساسان حسینعلی زاده[2]
[1]کارشناسی ارشد، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه آزاد قزوین، قزوین،
enosouhi@gmail.com
[2]استادیار، پژوهشکده فناوری اطلاعات، پژوهشگاه ارتباطات و فناوری اطلاعات، تهران،
s.alizadeh@itrc.ac.ir

چکیده

با توجه به اضافه‌بار اطلاعات در شبکه‌ی گسترده‌ی جهانی، کاربر در انتخاب اقلام با مشکل مواجه است. سرویس‌های فهرست بندی اجتماعی به کاربران اجازه می‌دهند که محصولات یا خدمات را استفاده کرده و نظرات و تجربیات خود را به اشتراک بگذارند که نه‌تنها برای خود، بلکه برای کاربران دیگر نیز مؤثر واقع می‌شوند. سیستم‌های توصیه‌گر با توجه به رفتار کاربران و ویژگی‌های محصولات، توانستند تأثیر بسزایی در انتخاب اقلام بگذارند. در این مقاله یک روش توصیه مبتنی بر برچسب، با توجه به احساسات منعکس‌شده در برچسب‌های کاربر پیشنهاد شده است. عملکرد روش پیشنهادی به این صورت است که امتیاز احساسی برچسب ها با دیگر اطلاعات دریافت شده از کاربر و همچنین اطلاعات محتوایی اقلام، ترکیب شده و میزان شباهت میان آن‌ها بدست آورده می‌شود. این فرآیند درنهایت، به بهبود عملکرد سیستم توصیه‌گر کمک می‌کند. آزمایش روش پیشنهادی بر روی پایگاه داده‌ی واقعی Movielense **نشان می‌دهد که این روش، عملکرد بهتری نسبت به کارهای قبلی دارد و در پیش‌بینی رتبه، خطا را کاهش و دقت را افزایش داده است.**

کلمات کلیدی

سیستم توصیه‌گر، برچسب، آنتولوژی، پیش‌بینی رتبه

۱- مقدمه

در عصر وب ۲، کاربران می‌توانند به تبادل و به اشتراک‌گذاری اطلاعات از طریق رسانه‌های آنلاین ازجمله وبلاگ‌ها، فیس‌بوک و انجمن‌ها بپردازند. درنتیجه مقدار زیادی گفتگو و اظهارنظر در شبکه‌ی گسترده‌ی جهانی تولید می‌شود. چنین اطلاعاتی می‌توانند جمع‌آوری و تجزیه‌وتحلیل شوند. به‌خصوص در تجارت الکترونیکی، سیستم توصیه‌گر، رفتارهای خرید قبلی کاربر و بازدیدهای آنلاین را برای کمک به کاربران در شناسایی سریع اقلام مناسب و یا موردعلاقه استفاده می‌کند[1].

همچنین با استفاده از منابع مختلف، افراد می‌توانند تجربیات و احساسات خود را پیوند دهند و از طریق رسانه های اجتماعی مختلف با دیگران در مورد علایق خود ارتباط برقرار کنند. برای آینده وب اجتماعی، سیستم‌های توصیه-گر اجتماعی با برچسب‌ها که در [3] معرفی شده، به‌طور گسترده به‌عنوان یک روش مهم برای بالا بردن کیفیت توصیه‌ها استفاده می‌شود.

سیستم‌های توصیه‌گر یک ابزار مفید برای پرداختن به بخشی از پدیده اضافه‌بار اطلاعات از اینترنت اثبات می‌شوند که تکامل آن‌ها با تکامل وب همراه است. سیستم‌های توصیه‌گر در [2] به ۳ دسته تقسیم می‌شوند:

۱- پالایش مشارکتی ۲- پالایش مبتنی بر محتوا ۳- پالایش ترکیبی.

در پالایش مشارکتی از اطلاعات گذشته کاربران برای پیش‌بینی اقلام موردعلاقه‌ی کاربر هدف استفاده می‌شود. در این روش به کاربر هدف اقلام موردعلاقه‌ی کاربران با سلیقه‌ی مشابه پیشنهاد می‌شود. روش‌های پالایش مشارکتی به دو گروه تقسیم می‌شوند: روش مبتنی بر حافظه و روش مبتنی بر مدل. در سیستم توصیه‌گر مبتنی بر محتوا، بر اطلاعات و ترجیحات کاربر تأکید می‌شود، بدین شکل که برای ارائه پیشنهاد، به انتخاب‌ها و تجربیات وی درگذشته توجه می‌شود.برای جلوگیری از محدودیت‌های خاص سیستم‌های پالایش مشارکتی و مبتنی بر محتوا، روش ترکیبی، روش مبتنی بر محتوا و پالایش مشارکتی را ترکیب می‌کند.

بااین‌حال سیستم‌های توصیه‌گر موجود، روابط اجتماعی میان کاربران را نادیده می‌گیرند و این نادیده گرفتن، باعث کاهش دقت توصیه به میزان قابل‌توجهی می‌شود[4].

سرویس‌های‌فهرست‌بندی‌اجتماعی(Goodreads، LibraryThing، Movielense و غیره) به کاربران اجازه می‌دهند که اقلام را فهرست بندی کرده و نظرات خود را با دیگران از طریق رتبه‌بندی، برچسب‌ها و نظرات به اشتراک بگذارند[2].

سیستم‌های توصیه‌گر متعارف، از اطلاعات رتبه‌بندی به‌عنوان بازخورد صریح کاربر در مورد اقلام استفاده کرده‌اند. برخلاف رتبه‌بندی، برچسب‌گذاری داده‌ها، به‌صراحت نشان‌دهنده‌ی اولویت کاربر برای محصول نیست، اما حاوی اطلاعات اضافی است. بنابراین، استفاده از داده‌های برچسب برای توصیه می‌تواند از تجربه کاربر پشتیبانی کند و اطلاعات رتبه‌بندی موجود را تکمیل نماید، درنتیجه امکان بهبود عملکرد توصیه را فراهم می‌کند[2].

مسئله این است که آیا استفاده از روابط معنایی میان برچسب‌ها (آنتولوژی) می‌تواند باعث بهبود دقت سیستم توصیه‌گر در پیش‌بینی‌ها و ارائه توصیه‌ها شود؟

هدف اصلی این پژوهش، افزایش و بهبود دقت سیستم توصیه‌گر در پیشنهاد اقلام به کاربران در شبکه‌های اجتماعی، میکروبلاگ ها، فروشگاه‌های آنلاین و سیستم فهرست بندی اجتماعی است.

ادامه این پژوهش این‌گونه دنبال خواهد شد: در بخش دوم، مروری بر تحقیقات انجام‌شده در این حوزه خواهد شد. بخش سوم در مورد روش پیشنهادی بحث خواهد شد. در بخش چهار نتایج بدست‌آمده در تحقیق ارزیابی‌شده و با روش‌های پیشین مقایسه می‌شود و بخش پنج شامل نتیجه‌گیری و پیشنهاد کار آتی در این راستا می‌باشد.

۲- مروری بر تحقیقات انجام‌شده

در این بخش مرور کوتاه بر تاریخچه سیستم توصیه‌گر کرده و سپس به بررسی پژوهش‌های انجام‌گرفته در خصوص سیستم توصیه‌گر مبتنی بر برچسب و آنتولوژی می‌پردازیم.

سیستم‌های توصیه‌گر معرفی شدند تا اقلام موردعلاقه‌ی کاربران را با پیش‌بینی علاقه‌ی کاربر به یک قلم بر اساس اطلاعات مرتبط در مورد اقلام، کاربران و تعاملات بین اقلام و کاربران، پیشنهاد کنند[5].

اولین مقاله پژوهشی در مورد سیستم‌های توصیه‌گر در اواسط دهه ۱۹۹۰ منتشر شد[6] و از آن زمان به بعد تحقیقات در این زمینه متنوع بوده و رویکردهای مختلف برای ارائه توصیه‌های بهتر معرفی‌شده اند. با گذشت زمان، استراتژی‌های جدیدتری در [7,8] از دسته‌های پایه با توصیه‌های بهبودیافته به وجود آمده که شامل اطلاعات اجتماعی، اطلاعات از اینترنت اشیا، اطلاعات مکان و روش های مبتنی بر الگوریتم ژنتیکی و غیره است. در دهه گذشته در هر دو زمینه صنعت و آموزشی، کارهای زیادی انجام شده است.

سیستم توصیه‌گر مبتنی بر برچسب، رفتار برچسب‌گذاری کاربر را برای تولید توصیه‌های مفید استفاده می‌کند. تفسیر کلمات کلیدی (برچسب‌ها) به‌عنوان یک راه جامع برای مرتبط کردن کاربر با اقلام و سپس ترکیب آن‌ها در سیستم‌های توصیه‌گر، یک گام امیدوارکننده برای افزایش کارایی و کیفیت توصیه‌ها در نظر گرفته می‌شود. برچسب‌ها می‌توانند سیستم توصیه‌گر را با اطلاعات اضافی غنی‌سازی کنند و ازاین‌رو، کیفیت توصیه‌ها را بهبود می‌بخشند. علاوه بر این، با استفاده از برچسب‌ها برای نشان دادن علائق کاربر، می‌توانیم توصیه‌هایی را برای سیستم‌هایی ارائه دهیم که اطلاعات رتبه‌بندی آن‌ها در دسترس نیست[9].

یکی دیگر از مزایای برچسب‌ها این واقعیت است که آن‌ها می‌توانند علائق در حال تغییر کاربر را در طول زمان ضبط کنند که این امر به‌سادگی می‌تواند با اضافه کردن برچسب‌های جدید به پروفایل کاربر، به‌روز شود. ازاین‌رو، شخصی‌سازی توصیه‌ها با توجه به علایق تغییر می‌کند[9].

در مدل ارائه‌شده به نام TNAM در [10]، یک ساختار مبتنی بر برچسب ارائه‌شده که اطلاعات برچسب و تعاملات محصول-کاربر را برای ارائه‌ی توصیه بر اساس علایق کاربر و ویژگی‌های محصول، ترکیب می‌کند. سیستم توصیه‌گری در [11] ارائه شده که از برچسب‌گذاری برای ارائه توصیه‌های مناسب در گروه‌های بحث و گفتگو استفاده می‌کند. برای این منظور، ارتباط معنایی برچسب‌ها با استفاده از پایگاه داده واژگان WordNet استخراج می‌شود و پرسش کاربر با استفاده از برچسب‌های معنایی مرتبط گسترش می‌یابد. کیم و همکارش در [12] یک روش توصیه، بر اساس روابط اعتماد ضمنی بدست آمده از اطلاعات برچسب گذاری کاربر پیشنهاد کردند. شوشین وی در [13] یک روش توصیه ترکیبی فیلم با استفاده از برچسب‌ها و رتبه‌بندی پیشنهاد می‌کند.

در علوم رایانه و علوم اطلاعات، آنتولوژی دربرگیرنده‌ی یک نماینده، نام‌گذاری رسمی و تعریف دسته‌ها، ویژگی‌ها و روابط بین مفاهیم، داده‌ها و اشخاص است که یک، چند یا همه‌ی دامنه‌ها را ماهیت می‌بخشد. تعریف اصلی از آنتولوژی در علوم رایانه توسط گروبر در سال ۱۹۹۲ ارائه شد و بعد توسط استاب و استودر در سال ۲۰۰۹ تصحیح شد. آنتولوژی معمولاً از واژگان و روابط بین مفاهیم تشکیل شده است. گانگ یک چارچوب سیستم توصیه‌گر قابل‌برنامه‌ریزی بر اساس داده‌های مرتبط با یکپارچه‌سازی داده‌های ارتباطی در حوزه آنتولوژی طراحی کرده است و از الگوریتم ژنتیک برای پردازش توصیه استفاده می‌کند[14].

لیم هیون در [2] یک روش توصیه مبتنی بر برچسب را با توجه به احساسات منعکس‌شده در برچسب‌های کاربر پیشنهاد می‌کند.

روش‌های توصیه‌ی ترکیبی جدیدی در [15] بر اساس روش‌های پالایش مشارکتی (CF) توسعه داده شده است. بر این اساس، دو ضعف اصلی سیستم‌های توصیه‌گر یعنی پراکندگی و مقیاس‌پذیری را با استفاده از کاهش ابعاد و روش‌های آنتولوژی حل می‌کند. در [16]، یک سیستم توصیه تبلیغات مبتنی بر آنتولوژی ارائه‌شده که از داده‌های تولیدشده توسط کاربران در

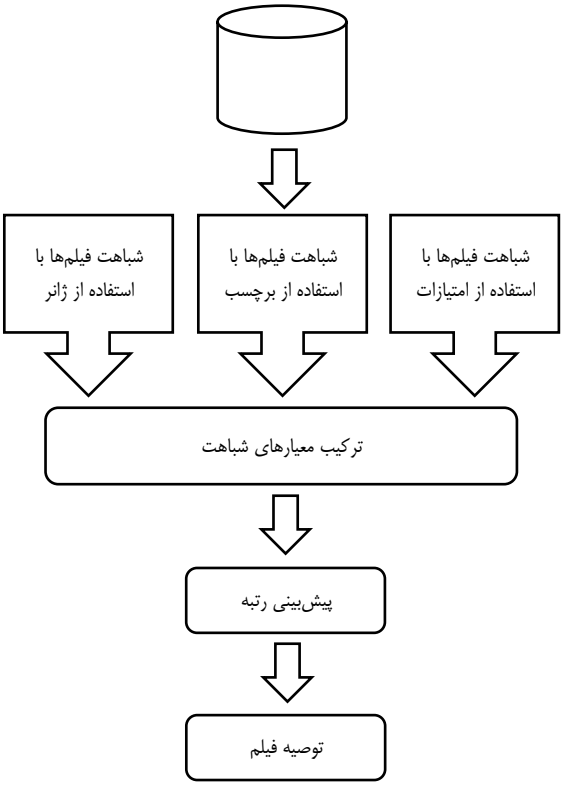
سایت‌های شبکه‌های اجتماعی استفاده می‌کند و این رویکرد با یک مدل آنتولوژی مشترک که نمایانگر پروفایل کاربران و محتوای تبلیغات است، اثبات می‌شود.

۳- روش پیشنهادی

شکل (۱)، شمای روش پیشنهادی را نشان می‌دهد. روش پیشنهادی ما با نام TagRS بیان می‌دارد که برای توصیه اقلام مناسب به کاربران (در پژوهش ما اقلام، فیلم است)، ابتدا باید شباهت میان این اقلام تعیین شود و سپس از طریق این شباهت‌ها توصیه مناسب به کاربر موردنظر داده شود.

جهت بررسی میزان شباهت، ما از سه روش، این شباهت را تعیین کرده و سپس این روش‌ها را با یکدیگر ترکیب کرده و نتیجه نهایی آن، شباهت کل میان دو فیلم را نشان می‌دهد. این روش‌ها عبارت‌اند از:

۱. بدست آوردن شباهت فیلم‌ها با استفاده از امتیازات داده‌شده توسط کاربران.
۲. بدست آوردن شباهت فیلم‌ها با استفاده ژانر آن‌ها.
۳. بدست آوردن شباهت فیلم‌ها با استفاده از برچسب‌هایی که کاربران به آن‌ها زده‌اند.



شکل (۱): شمای کلی روش پیشنهادی

علت انتخاب این روش‌ها این است که بدست آوردن شباهت اقلام با استفاده از رتبه‌های کاربران و نوع اقلام در پژوهش‌های قبلی، اگرچه با موفقیت روبرو بوده‌اند و توانسته‌اند عملکرد خوبی از خود نشان دهند ولی با ترکیب موارد مذکور با برچسب‌ها به‌عنوان روشی نوین و استفاده از اطلاعات احساسی آن‌ها، می‌تواند باعث اثربخشی بیشتر و خطای کمتر در پیش‌بینی رتبه‌ها و سیستم توصیه گر شود.

۳-۱- شباهت اقلام با استفاده از امتیازات داده‌شده توسط کاربران

هر کالا در سیستم می‌تواند توسط کاربران مختلف امتیازدهی شده باشد و هر کاربر می‌تواند برخی اقلام را امتیاز داده باشد و برخی دیگر بدون امتیاز باشند. در سیستم Movielens[۱]، محدوده‌ی این امتیازها از ۰ الی ۵ است.کاربری که فیلمی را امتیاز ۵ داده باشد به این معناست که این فیلم موردعلاقه‌ی کاربر است و هر چه این عدد به ۰ نزدیک‌تر شود، نشان‌دهنده‌ی عدم علاقه‌ی کاربر به فیلم است. علت انتخاب این معیار به‌عنوان یکی از سه معیار روش پیشنهادی ما، این است که همیشه رتبه‌ها اطلاعات مفیدی برای کاربران محسوب می‌شوند و استفاده از آن‌ها باعث می‌شود که کاربران ترغیب به استفاده از فیلم‌های با رتبه‌ی بالاتر شوند. ما برای بدست آوردن شباهت دو فیلم با استفاده از امتیازها، از معیار شباهت فاصله‌ی کسینوسی طبق فرمول (۱) استفاده می‌کنیم. شباهت کسینوسی بیان می‌دارد:

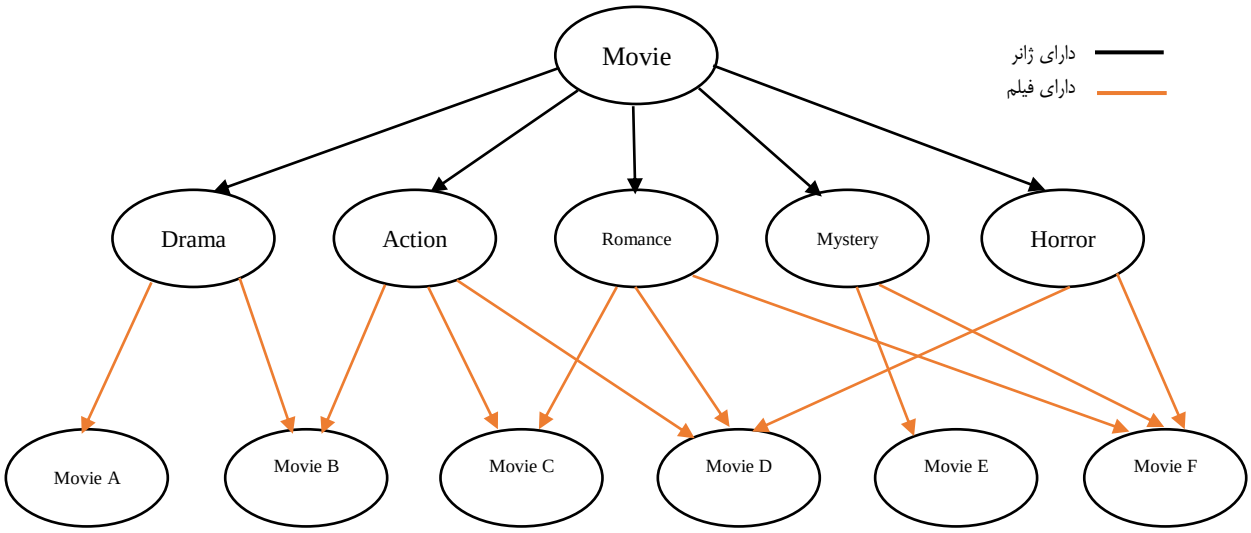
- اگر دو بردار (متن) یکسان باشد، مقدار فاصله کسینوسی برابر است با ۱.
- اگر دو بردار (متن) کاملاً متفاوت باشد، مقدار فاصله کسینوسی برابر است با ۰.

$$\mathrm{Cos}(\Theta) = \frac{A.B}{\|A\|.\|B\|} = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2}\sqrt{\sum_{i=1}^{N} B_i^2}} \qquad (۱)$$

مقدار، عددی بین ۰ و ۱ است. در خصوص شباهت میان دو فیلم از طریق این روش، هرچه عدد بدست‌آمده به ۱ نزدیک‌تر باشد، میزان شباهت بیشتر است.

۳-۲- شباهت اقلام با استفاده از محتوا

ما شباهت معنایی فیلم‌ها را از طریق یک معیار وابستگی معنایی مبتنی بر درخت سلسله مراتبی محاسبه می‌کنیم. رویکرد مبتنی بر درخت می‌تواند شامل کل دامنه آنتولوژی باشد؛ یک نمونه از این درخت در شکل(۲) نشان داده شده است. درخت فیلم با یک گره «فیلم» شروع می‌شود که با شاخه‌های مختلف ژانر (درام، اکشن، عاشقانه،رمز و رازی و ترسناک) به‌وسیله‌ی رابط "دارای ژانر" وصل می‌شود. پس از آن، فیلم‌ها با رابط "دارای فیلم" به ژانر مربوطه



شکل(۲): مثالی از ساختار درختی برای آنتولوژی فیلم

متصل می‌شوند. لازم به ذکر است که یک فیلم می‌تواند دارای انواع مختلف ژانر باشد و بنابراین می‌تواند به شاخه‌های مختلف ژانر متصل شود.

بااین‌حال، برای سادگی، ما فقط از آنتولوژی MO [17] [i i] در این رویکرد مبتنی بر درخت استفاده می‌کنیم. MO زبان کنترل شده‌ای برای توضیح اصول و مفاهیم معنایی مرتبط با حوزه‌های فیلم فراهم می‌کند.

ما برای این کار از شباهت معنایی بانام SemSim که در [18] پیشنهادشده، استفاده می‌کنیم، زیرا این روش به‌درستی به بررسی محتوای اقلام (در اینجا فیلم‌ها) می‌پردازد و در ارائه نتیجه‌ی این فرآیند، کمک دهنده خواهد بود. روال این روش به این صورت است که شباهت بین هر جفت فیلم به‌وسیله‌ی شبه کد محاسبه می‌شود. همان‌طور که در الگوریتم (۱) نشان داده‌شده است، درصد تشابه معنایی بین یک جفت فیلم (هر فیلم به‌عنوان یک بردار دودویی است) از طریق ضرایب تشابه ژاکارد دودویی در فرمول (۲) محاسبه می‌شود. با استفاده از ضریب ژاکارد می‌توان میزان صفت‌های مشترک بین دو شئ (در اینجا اشیاء، فیلم‌ها و صفت موردبررسی، ژانر آن‌ها است) را محاسبه کرد.

$V_i = (v_{i,1}, v_{i,2}, \ldots, v_{i,g})$

$$v_{i,g} = \begin{cases} ۱ & \text{اگر فیلم } i \text{ ژانر } g \text{ داشته باشد} \\ ۰ & \text{اگر فیلم } i \text{ ژانر } g \text{ نداشته باشد} \end{cases}$$

V_i ارزش بردار فیلم i است و g نوع ژانر فیلم است.

در این تعریف، اگر i و j فیلم‌های ما باشند:

$$SemSim = \frac{Q_{11}}{(Q_{01}+Q_{10}+Q_{11})} \qquad (۲)$$

Algorithm1:Semantic Similarity Calculation[18]

//**GenreMat** is a *n*m* matrix
// *n* is the number of movies and *m* is the number of genre.
Require GenreMat,n
//calculate semantic similarity for all pair of movies.
For M_i from 1 to n **do**
  For M_j from 1 to n **do**
    $Q01 \leftarrow$ number of cells where M_i is 0 and M_j is 1.
    $Q10 \leftarrow$ number of cells where M_i is 1 and M_j is 0.
    $Q11 \leftarrow$ number of cells where M_i is 1 and M_j is 1.
      If ($Q01+Q10+Q11$) is 0 **then**
        SemSim(M_i,M_j)$\leftarrow$0
      Else
        SemSim(M_i,M_j)$\leftarrow Q11/ (Q01+Q10+Q11)$
      End
  End for
End for
Return SemSim for all pair of (M_i,M_j)

$Q11$: تعداد کل ژانرهایی که هم I و هم J دارند ($v_{i,g}$و $v_{j,g}$ برابر ۱ است).

$Q01$: تعداد کل ژانرهایی که I ندارد ولی J دارد ($v_{i,g}$ برابر ۰ و $v_{j,g}$ برابر ۱ است).

$Q10$: تعداد کل ژانرهایی که J ندارد ولی I دارد ($v_{i,g}$ برابر ۱ و $v_{j,g}$ برابر ۰ است).

با توجه به این توضیحات، الگوریتم SemSim به این صورت تعریف می‌شود:

در این الگوریتم منظور از Mi و Mj، همان Movie i و Movie j است. اگر مجموع Q01،Q11 و Q10 برابر با ۰ شود، میزان شباهت دو فیلم هم ۰ حساب می‌شود. در غیر این صورت میزان شباهت میان دو فیلم از فرمول ۲ محاسبه می‌گردد. این عملیات برای همه‌ی جفت فیلم‌های موجود در پایگاه داده تکرار می‌گردد.

۳-۳- شباهت اقلام با استفاده از برچسب‌ها

در وب‌سایت Movielense، کاربران علاوه بر دادن امتیاز، می‌توانند برچسب‌هایی را نیز به هر فیلم بزنند. هر برچسب می‌تواند حاوی اطلاعات مفیدی راجع به عقیده یا احساسات کاربر نسبت به فیلم‌ها باشد. این احساسات می‌توانند مثبت یا منفی باشند.کاربر می‌تواند به هر فیلم چندین برچسب بزند و برای هر فیلم می‌تواند توسط کاربران مختلف برچسب‌های متعددی درج شده باشد.

در این پژوهش، جهت تخمین احساسات کاربران نسبت به فیلم‌های مختلف، از فرهنگ لغت SenticNet [iii] استفاده شده است که این فرهنگ لغت شامل ۱۰۰۰۰۰ کلمه و عبارت با نشان دادن میزان احساسات مخفی درون آن‌ها است. در این فرهنگ لغت احساسات به‌صورت عددی بین [۱-و ۱+] نشان داده می‌شوند. اگر کلمه‌ای دارای احساسات یا بار منفی باشد، عدد درج‌شده منفی است و اگر کلمه دارای احساسات یا بار مثبت باشد عدد درج‌شده مثبت است. هرچه به ۱- نزدیک‌تر باشد میزان احساسات منفی بیشتر و هرچه این عدد به ۱+ نزدیک‌تر باشد نشان‌دهنده بیشترین احساسات مثبت است. به‌عنوان‌مثال برای کلمه sad، عدد ۰.۹۱- و برای کلمه funny، ۰.۹۲۸ درج شده است. علت انتخاب این دیکشنری در میان دیگر نمونه‌های مشابه، در دسترس بودن آن و اطلاعات دقیق عددی آن است.

فرمول (۳) بیان می‌کند وزن احساسی هر برچسب، همان امتیاز احساسی برچسب موجود در فرهنگ لغت SenticNet است.

$$Weight_{emotion}(t_{u,i}) := EmotionScore(t_{u,i}) \qquad (۳)$$

(در این معادله برای تعریف، از نماد := استفاده شده است.)

برای بدست آوردن شباهت دو فیلم از طریق برچسب‌های زده شده توسط کاربر، همان‌طور که در [2] بیان شده، ابتدا عملیات پیش‌پردازش روی مجموعه داده باید انجام گیرد. عملیات انجام‌شده به این صورت است:

- برچسب‌هایی که حاوی اسامی خاص (مثل اسم بازیگر و غیره) حذف‌شده‌اند.
- برچسب‌های حاوی علائم خاص (مثل !، ؟، * و غیره) حذف‌شده‌اند.
- برچسب‌هایی که در فرهنگ لغت یافت نشده‌اند، حذف‌شده‌اند.
- اگر فیلمی دارای چند برچسب باشد، برای بدست آوردن میزان احساسات نهایی کاربر بر فیلم، با جایگذاری مقدار احساسات هرکدام از کلمات در فرمول (۴)، مقدار احساسات کل برچسب به‌صورت میانگین محاسبه‌شده و این عدد برای تک‌تک کلمات برچسب درج شده است.

$$Weight(t_{u,i}) = \frac{1}{|t_{u,i}|}\sum_{k=1}^{|t_{u,i}|} Weight(t_{u,i}^{k}) \qquad (۴)$$

که $t_{u.i}^{k}$، kامین برچسبی است که کاربر u به فیلم i زده است.

پس از انجام عملیات پیش‌پردازش روی پایگاه داده، شباهت میان دو فیلم از طریق فرمول (۵) محاسبه می‌شود:

$$Sim_{Tag} = 1 - \frac{|a-b|}{Max} \qquad (۵)$$

که a و b وزن کل برچسب زده‌شده به دو فیلم و Max بیشترین فاصله میان وزن برچسب‌ها در کل مجموعه‌ی داده است (به‌طور پیش‌فرض ۱.۹۱۸).

نتیجه عددی بین ۰ و ۱ خواهد بود که هرچه عدد بدست‌آمده به یک نزدیک‌تر باشد، میزان شباهت دو فیلم نیز بیشتر است.

پس‌ازاینکه شباهت فیلم‌ها از سه روش مبتنی بر رتبه،ژانر و برچسب بدست آمد، شباهت نهایی دو فیلم از طریق فرمول(۶) بدست می‌آید:

$$Similarity = \frac{Sim_{Rating} + SemSim + Sim_{Tag}}{3} \qquad (۶)$$

عدد بدست‌آمده بین ۰ و ۱ خواهد بود. هرچه به ۱ نزدیک‌تر، شباهت دو فیلم بیشتر است. به‌عنوان مثال، دو فیلم از طریق برچسب‌ها میزان شباهت ٪۶۵ از طریق SemSim ٪۸۵ و از طریق رتبه‌بندی‌ها ٪۹۰ بدست می‌آورند. درنتیجه، شباهت کل این دو فیلم با استفاده از فرمول ۶ مقدار ٪۸۰ بدست می‌آید.

۳-۴- پیش‌بینی رتبه

فرمولی که برای پیش‌بینی رتبه فیلم، مورداستفاده قرارگرفته است به شکل فرمول (۷) زیر بیان می‌گردد:

$$\hat{R}_{u,T_i} = \frac{\sum_j S_{T_{i,j}} . R_{u,j}}{\sum_j S_{T_{i,j}}} \qquad (۷)$$

T_i، فیلم هدف است که می‌خواهیم رتبه‌ی آن را پیش‌بینی کنیم، j سایر فیلم‌هایی که کاربر u به آن‌ها رتبه داده است، $S_{T_{i,j}}$ مقدار عددی شباهت فیلم هدف و سایر فیلم‌ها است و $R_{u,j}$ رتبه‌ای که کاربر u به بقیه-ی فیلم‌ها داده است. $S_{T_{i,j}}$ با استفاده از سه روش بیان‌شده در پژوهش،

محاسبه شده و با $R_{u,j}$ ترکیب می‌شود. درنتیجه‌ی این فرآیند، $\hat{R}_{u,T_i}$، که بیانگر رتبه پیش‌بینی‌شده برای فیلم هدف است، بدست می‌آید.

برای این کار، کاربر شناسه‌ی فیلم موردنظر را وارد می‌کند و سیستم با توجه به سابقه‌ی کاربر و محاسبه‌ی شباهت این فیلم با فیلم‌های رتبه‌بندی شده توسط کاربر در سیستم، رتبه‌ی فیلم مذکور را محاسبه می‌کند.

۳-۵- توصیه فیلم(های) برتر

هر کاربری که در سیستم عضو باشد ممکن است سابقه‌ای از خود به‌صورت دادن امتیاز و زدن برچسب به فیلم‌ها، در سیستم به‌جا گذاشته باشد. از برچسب‌ها و رتبه‌ها و ژانر ثبت‌شده برای فیلم‌ها، قبلاً شباهت میان فیلم‌ها محاسبه شده است. حال برای دادن توصیه مناسب به کاربر u، فقط از رتبه‌ها استفاده می‌کنیم و فیلم‌های موردعلاقه‌ی کاربر را به او پیشنهاد می‌کنیم. اگر کاربر u به فیلم i، رتبه‌ی بیشتر از ۳ داده باشد می‌توان فرض کرد که فیلم i، فیلم موردعلاقه‌ی کاربر u است. با توجه به این علاقه‌مندی می‌توان فیلم‌های مشابه با فیلم i را پیدا کرد و آن‌ها را به کاربر u پیشنهاد داد. هرچه میزان این شباهت بیشتر باشد، اولویت برای توصیه بالاتر می‌رود.

برای این کار کافی است کاربر، کد کاربری خود و تعداد پیشنهاد موردنیاز را وارد سیستم کند تا سیستم به او پیشنهادهای مناسب را ارائه کند.

٤- ارزیابی

ما از پایگاه داده‌ی Movielens که یک سرویس فهرست بندی اجتماعی فیلم است، برای ارزیابی عملکرد روش پیشنهادی استفاده می‌کنیم. مجموعه‌ی داده دارای ۷۱۵۶۷ کاربر، ۱۰۶۸۱ فیلم، ۱۰۰۰۰۰۵۴ رتبه‌بندی و ۹۵۵۸۰ سابقه برچسب گذاری است. ۱۵۲۳۰ برچسب متمایز و ۴۰۰۹ کاربر که برچسب‌ها را حداقل یک‌بار استفاده می‌کنند، وجود دارد. به‌طور متوسط هر کاربر ۱۴۳ فیلم رتبه‌بندی می‌کند و هر فیلم دارای ۱۰ برچسب مجزا است. در میان کاربرانی که سابقه برچسب گذاری دارند، ۴۰ درصد از کاربران تنها از یک برچسب استفاده می‌کنند.

ما داده‌ها را برای آزمایش به کاربران و فیلم‌ها با سابقه‌ی برچسب‌گذاری محدود کردیم و تعدادی از داده‌های بدون استفاده نیز از پایگاه داده حذف شد، که درنهایت پایگاه داده مورداستفاده شامل ۶۷۱ کاربر، ۱۰۰۰۰۵ رتبه، ۹۱۲۵ فیلم و ۴۰۴ برچسب می‌باشد.

برای انجام ارزیابی، به دلیل تعداد زیاد رتبه‌ها، ما از روش Holdout [19] استفاده کرده و پایگاه داده را به دو قسمت پایگاه داده آموزشی (۸۰٪) و پایگاه داده آزمون (۲۰٪) تقسیم کرده و ارزیابی را روی پایگاه داده آزمون اجرا کرده‌ایم. ما برای ارزیابی، رتبه‌ی داده‌شده به یک فیلم توسط کاربر را با استفاده از آنتولوژی، پیش‌بینی کرده و سپس با مقدار رتبه‌ای که کاربر به فیلم داده است، مقایسه می‌کنیم و با روش‌های ارزیابی میزان دقت و کیفیت پیش‌بینی رتبه را بدست می‌آوریم.

معیارهای ارزیابی میانگین خطای مطلق یا MAE (۸)[18]، F-measure (۹) و دقت (۱۰)[20] در این پژوهش محاسبه شده است.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|R_p - R_i| \quad (۸)$$

$$F - Measure = \frac{2 * Precision * RC}{Precision + RC} \quad (۹)$$

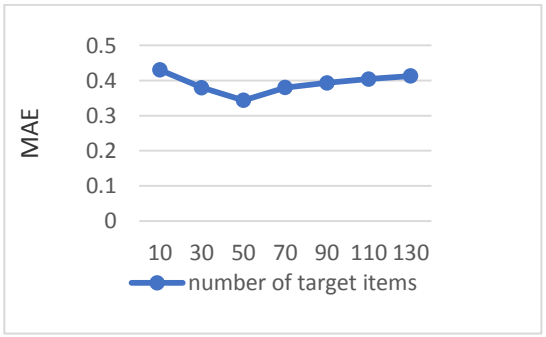
$$Precision = 1 - \frac{MAE}{R_{Max} - R_{Min}} \quad (۱۰)$$

R_P رتبه پیش‌بینی‌شده، R_i رتبه واقعی، R_{Max} بیشترین رتبه در پایگاه داده آزمون، R_{Min} کمترین رتبه در پایگاه داده آزمون، N تعداد کل رتبه‌ها در پایگاه داده‌ی آزمون و RC فاکتور پوشش تعریف‌شده‌اند.

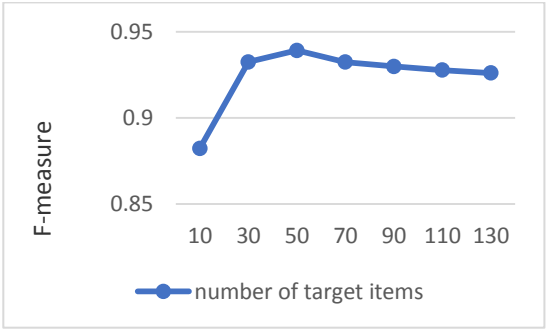
ارزیابی‌ها، روی ۲۰٪ از مجموعه‌ی داده Movielense به‌عنوان مجموعه داده‌ی آزمون، اجرا شده است. مقدار MAE هرچه کمتر باشد بهتر است و مقدار F-measure هر چه بیشتر باشد بهتر است. مقادیر MAE و F-measure برای تعداد اقلام هدف (Target items) مختلف (N=۱۰،۳۰،۵۰،۷۰،۹۰،۱۱۰،۱۳۰) بررسی‌شده و نتایج آن در شکل (۳) و (۴) نشان داده شده است. محور X، Ti ها و محور Y محدوده‌ی مقادیر MAE و F-measure را نمایش می‌دهد.

مقدار MAE در همه‌ی حالات بسیار نزدیک به یکدیگر است ولی در N=۵۰ در کمترین و بهترین حالت خود قرار دارد و نشان‌دهنده‌ی کمترین میزان خطا است. مقدار F-measure نیز در حالت N=۵۰ در بیشترین و بهترین حالت خود قرار دارد. درواقع استفاده از داده‌های برچسب باعث بهبود روش‌های قبلی شده است. در آزمایش‌های صورت گرفته بر روی روش‌های مبتنی بر رتبه و مبتنی بر محتوا، در محاسبه ی میزان شباهت فیلم ها و همینطور پیش بینی رتبه، چنین نتیجه ای حاصل نشد. ولی با اضافه کردن داده‌های برچسب و ترکیب اطلاعات این سه روش، خطای کمتر و دقت بیشتر نتیجه‌ی اثربخشی این روش را تایید کرد.

درنهایت مقادیر MAE روش پیشنهادی که با نام TagRS تعریف‌شده، با نتایج مقالات UBBCF و UBHUS (که دو روش CF مبتنی بر کاربر هستند) [21,22]، SoTrust (روشی که می‌تواند رتبه‌ها را برای سیستم توصیه‌گر سفارشی، بر اساس شباهت، مرکزیت و روابط اجتماعی پیش‌بینی کند) [23]، FPMF (که معیار شباهت را به‌وسیله‌ی ارتباطات مستقیم و غیرمستقیم میان کاربران را دریافت کرده و با اطلاعات رتبه‌بندی آن‌ها ترکیب کرده و از آن‌ها برای پیش‌بینی رتبه استفاده می



شکل (۳): نمودار MAE برای تعداد N مختلف



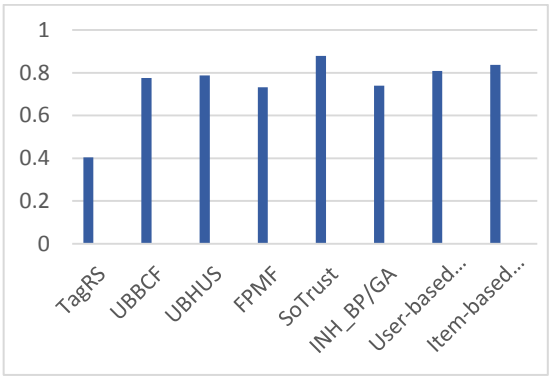
شکل (٤): نمودار F-measure برای تعداد N مختلف

کند) [24]، دو پالایش مشارکتی مبتنی بر کاربر و قلم [25]،INH-BP/GA (روشی که امکان شخصی‌سازی پیش‌بینی را متناسب با علایق کاربر فراهم می‌کند و از یک الگوریتم بهینه‌سازی مانند الگوریتم ژنتیک استفاده می‌کند) [26]، در نمودار شکل (۵) مقایسه شده‌اند.

روش پیشنهادی با استفاده از تعیین میزان شباهت‌ها میان اقلام توانسته کمترین میزان خطا و بیشترین دقت در پیش‌بینی رتبه‌ی کالای هدف (T_i) نسبت به روش‌های دیگر، از خود نشان دهد. این افزایش دقت به‌نوبه‌ی خود باعث بهبود دقت در نتایج پیشنهادهای سیستم توصیه‌گر نیز خواهد شد.

۵- نتیجه‌گیری

تاکنون روش‌های مختلفی برای توصیه‌ی اقلام در وب‌سایت‌های تجارت الکترونیک، شبکه‌های اجتماعی و سرویس‌های فهرست بندی اجتماعی که به‌طور تخصصی در حوزه توصیه فیلم، موسیقی، کتاب و غیره فعالیت دارند، به کار گرفته شده است. کاربران در سرویس فهرست بندی اجتماعی، اقلام را فهرست بندی کرده و تجربیات خود را با دیگران به اشتراک می‌گذارند . بارگذاری مطالب مختلف باعث می‌شود کاربران در انتخاب اقلام دچار مشکل شوند. سیستم توصیه‌گر، مشکل انتخاب را با توصیه قلم با توجه به رفتار کاربر و ویژگی‌های محتوا، کاهش می‌دهد.



شکل (۵): مقایسه MAE روش پیشنهادی با پژوهش‌های قبلی

در این پژوهش ما یک روش مبتنی بر برچسب را با توجه به احساسات منعکس‌شده در برچسب‌های کاربر و ترکیب آن با رتبه‌های داده‌شده به اقلام و نوع محتوای فیلم‌ها، پیشنهاد کرده‌ایم. برآورد کاربر از این اقلام پس از مصرف اقلام، ایجادشده است؛ بنابراین، احساسات کاربر به طور مستقیم منعکس می‌شوند و می توانند نقش مهمی را در سیستم توصیه‌گر بازی کنند. رتبه‌بندی‌ها ارزش مثبت یا منفی اقلام را نشان می‌دهند و برچسب‌ها دلیل دقیق‌تری برای برآورد است. بنابراین، هنگامی‌که یک کاربر یک قلم را رتبه‌بندی کرده و برچسب‌گذاری کرده است، از امتیاز قلم به عنوان احساس پایه برچسب استفاده می‌کنیم و وزن برچسب داده‌شده به فیلم را بر اساس SenticNet که فرهنگ لغت احساسی است، تنظیم می‌کنیم. سپس با ترکیب این مقادیر، میزان شباهت کلی میان اقلام را بدست می‌آوریم که از این شباهت برای پیش‌بینی رتبه در سیستم توصیه‌گر استفاده می‌کنیم.

استفاده از این روش به دلیل انتخاب صحیح و استفاده از روش‌های مناسب و جدید در حیطه‌ی پژوهش، باعث افزایش دقت در پیش‌بینی رتبه نسبت به پژوهش‌های انجام گرفته‌ی قبلی شده، که این دستاورد باعث افزایش دقت در عملکرد سیستم توصیه گر خواهد شد. درنتیجه، مسئله این پژوهش را توانست به‌درستی پوشش دهد و نشان دادیم که استفاده از روابط معنایی میان برچسب‌ها در سیستم توصیه گر، می‌تواند با درصد خطای کم، کمک دهنده و مفید واقع شود.

عدم دستیابی به مجموعه‌ی داده جامع و بزرگ از بزرگ‌ترین محدودیت‌های این پژوهش محسوب می‌شود. به‌ویژه برای پایگاه داده برچسب‌های زده‌شده به فیلم‌ها که پایگاه داده نسبتاً کوچک بود و برای بسیاری از فیلم‌ها برچسبی تعریف نشده بود. همچنین در مجموعه داده‌ها پراکندگی وجود داشت و برای بسیاری از اقلام رتبه‌ای تعریف نشده بود. همچنین در پژوهش حاضر، مجموعه داده‌ی حاوی برچسب‌ها، علاوه بر لغات و عبارات، شامل علائم خاص (نظیر !، ؟، * و غیره) نیز بود که میزان احساسات کلمات و عبارات حاوی این علائم غیرقابل‌بررسی بود و حذف شدند. به عبارتی اگر لغتی حاوی ! باشد، مثبت یا منفی بودن احساسات

کـاربر، غیرقابـل‌بررسـی اسـت زیـرا چنـین علائمـی در فرهنـگ لغـت SenticNet تعریف نشده است.

به‌عنوان کار آتی می‌توان از فرهنگ لغـت دیگـری کـه بـرای ایـن علائم مقدار عددی منحصربه‌فردی با نشان دادن مثبت یا منفی بودن کلمه حاوی این علائم، ارائه داده است، استفاده کرد. همچنین مـی‌تـوان میـزان احساسات این علائم را با روش‌های ریاضی یا آماری تعریف کرد تـا دقـت در میزان شباهت میان فیلم‌ها افزایش یابد. همچنین می‌توان بـرای ایجـاد شباهت میان برچسب‌های زده‌شده به فیلم‌ها، از ایجاد رابطـه بـه‌وسـیله‌ی مترادف‌ها و متضادها میان لغات استفاده کرد.

مراجع

[1] Huang, T. C. K., Chen, Y. L., & Chen, M. C. A novel recommendation model with Google similarity. Decision Support Systems, 89:pp.17–27, 2016.

[2] Lim, H., & Kim, H. J. Item recommendation using tag emotion in social cataloging services. Expert Systems with Applications, 89, 179–187, 2017.

[3] Heung-NamKim,Abdulmajeed Alkhaldi,Abdulmotaleb El Saddik,Geun-Sik Jo. Collaborative user modeling with user-generated tags for social network. Expert Systems with Applications,pp.8488–8496, 2011.

[4] Young-Duk Seo, Young-Gab Kim, Euijong Lee, Doo-Kwon Baik. Personalized Recommender System based on Friendship Strength in Social Network Services. Expert Systems With Applications,pp.135-148, 2016.

[5] Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. Recommender systems survey. Knowledge-Based Systems,46:pp.109–132, 2013.

[6] Shardanand, U., & Maes, P. Social information filtering:pp. 210–217,1995.

[7] Burke R. Hybrid recommender systems: survey and experiments. User Model User-adapted Interact; 12(4):pp.331–70,2002.

[8] M. Pazzani.A framework for collaborative, content-based, and demographic filtering, Artificial Intelligence Review-Special Issue on Data Mining on the Internet 13 (5-6):pp.393–408,1999.

[9] Schroeder, U., Chatti, M. A., Dakova, S., & Thu, H. Tag-Based Collaborative Filtering Recommendation in Personal Learning Environments, 6(4),pp. 337–349, 2013.

[10] Ruoran Huang, N. W. TNAM: A tag-aware neural attention model for Top-N recommendation. Neurocomputing, 1-12, 2019.

[11] Masoumeh Riyahi, M. K. Providing effective recommendations in discussion groups using a new hybrid. Electronic Commerce Research and Applications, 1-24, 2020.

[12] Kim, H., & Kim, H.-J.Improving recommendation based on implicit trust re-lationships from tags. In Proceedings of the 2nd international conference on com-puters, networks, systems, and industrial applications, pp. 25–30, 2012.

[13] Shouxian Wei, X. Z. A Hybrid Approach for Movie Recommendation via Tags and Ratings. Electronic Commerce Research and Applications, pp.83-94, 2016.

[14] Rui Ren, L. Z. Personalized Financial News Recommendation Algorithm Based on Ontology. Procedia Computer Science,pp.843-851, 2015.

[15] Mehrbakhsh Nilashi, O. I. A. Recommender System Based on Collaborative Filtering Using Ontology and Dimensionality Reduction Techniques. Expert Systems With Applications,pp. 507- 520, 2017.

[16] Francisco García-Sáncheza, R. C.-P.-G.A social-semantic recommender system for advertisements. Information Processing and Management, 1-16,2020.

[17] Bouza, A.,MO–the Movie Ontology, 2010.

[18] Ranjbar Kermany, N., & Alizadeh, S. H. A hybrid multi-criteria recommender system using ontology and neuro-fuzzy techniques. Electronic Commerce Research and Applications, 21:pp. 50–64, 2017.

[19] Kohavi, R A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI,1995.*

[20] J. Bobadilla, F. O.Recommender systems survey. Knowledge-Based Systems, 109-132, 2013.

[21] Bidyut Kr. Patra, R. L.A new similarity measure using Bhattacharyya coefficient. Knowledge-Based Systems, 163-177, 2015.

[22] Yong Wang, J. D. A hybrid user similarity model for collaborative filtering. Information Sciences, 102-118, 2017.

[23] Anahita Davoudi, M. C. Social trust model for rating prediction in recommender systems: Effects of similarity, centrality, and social ties. Online Social Networks and Media, 1-11, 2018.

[24] Chenjiao Feng, J. L. A fusion collaborative filtering method for sparse data in recommender systems. Information Sciences, 365-379, 2020.

[25] Sarwar, B., Karypis, G., Konstan, J., and Reidl, J. Itembased collaborative filtering recommendation algorithms. In Proceedings of the 10th international Conference on World Wide Web (Hong Kong, Hong Kong, May 01 - 05, 2001). ACM, New York, NY,pp. 285-295, 2001

[26] Bushra Alhijawi, G. A.-N. Novel predictive model to improve the accuracy of collaborative filtering. Information Systems, 1-33, 2020

زیرنویس

[i] https://grouplens.org/datasets/movielens/

[ii] http://www.movieontology.org

[iii] https://sentic.net/