سال چهاردهم، شماره‌های ۵۱ و ۵۲ ، بهار و تابستان ۱۴۰۱
صص: ۷۳_۸۲

دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

سیستم توصیه گر فیلم فیلتر اشتراکی مبتنی بر ضریب همبستگی بین کاربران و محاسبه میانگین وزنی امتیازات با دقت بالا

نوذر ابراهیمی لامع* دکتر فاطمه ثقفی دکتر مجید قلی پور*****

*دانشجوی دکترای مدیریت فناوری اطلاعات، گروه مدیریت سیستم های اطلاعاتی، دانشگاه آزاد اسلامی، واحد قزوین، قزوین، ایران
**دانشیار دانشکده مدیریت دانشگاه تهران، تهران، ایران
***استادیار دانشکده برق و رایانه،واحد قزوین، دانشگاه آزاد اسلامی ، قزوین، ایران
تاریخ دریافت:۱۴۰۰/۰۱/۱۴ تاریخ پذیرش:۱۴۰۰/۰۶/۰۹
نوع مقاله: پژوهشی

چکیده

سیستم های توصیه گر، وظیفه راهنمایی و هدایت کاربر جهت انتخاب بهینه آیتم ها، مطابق با علائق و سلیقه های آنها را بر عهده دارند. علی رغم حدود سه دهه سابقه تحقیقات بر روی سیستم های توصیه گر، ولی موضوع مذکور هنوز یکی از چالش های تحقیقاتی به روز می باشد. این سیستم ها با شخصی سازی پیشنهادات، باعث صرفه جویی در وقت و افزایش رضایت کاربران می گردند. این سیستم ها در اغلب سایت های معتبر خارجی و داخلی مورد استفاده قرار گرفته اند. در سیستم های توصیه گر، مهم ترین و پرکاربردترین روش پالایش داده ها، روش پالایش اشتراکی می باشد. در این مقاله نسبت به پیاده سازی سه سیستم توصیه گر فیلتر اشتراکی مبتنی بر محاسبه ضریب همبستگی بین کاربران، انتخاب تعداد بهینه همسایه ها و محاسبه امتیازات وزنی اقدام شده و بهترین روش با کمترین خطا به عنوان مدل مورد نظر انتخاب شده است. ورودی سیستم داده های تحقیقاتی مووی لنز با حدود ۱۰۰ هزار امتیاز می باشد. روش بکار رفته نسبت به آخرین مقاله ای که از روش همبستگی ترکیبی استفاده کرده است ۳/۲۹ درصد مقدار خطای RMSE را بهبود می بخشد.

واژگان کلیدی: سیستم های توصیه گر، ضریب همبستگی، فیلتر اشتراکی، فیلم.

۱. مقدمه

آغاز عصر اینترنت و ارتقاء آن به وب ۲[۱]و امکان تبادل اطلاعات دو طرفه، باعث تغییرات وسیعی در بسیاری از عرصه ها از جمله شروع و گسترش روز افزون تجارت الکترونیک[۲] گردید[۱][۲].

در این بین امکان ارائه پیشنهادات شخصی سازی شده برای هر کاربر بصورت اختصاصی باعث توسعه بیشتر این نوع تجارت خصوصا" در شاخه B2C گردید. زیرا شخصی سازی پیشنهادات، ضمن صرفه جویی در زمان جستجو و افزایش رضایت خریدار، باعث گردش سریع تر کسب وکار و افزایش فروش می گردد[۲][۳][۴] [۵].

نویسنده مسئول: دکتر فاطمه ثقفی fsaghafi@ut.ac.ir

۱ Web 2

۲ Electronic Commerce/e-commerce

برای ارائه چنین توصیه های اختصاصی، نیاز به سیستم های نرم افزاری می باشد که از اطلاعات آیتم ها و کاربران استفاده کرده و با الگوریتم ها و روش های مختلف داده کاوی، نسبت به پردازش اطلاعات و ارائه پیشنهادات متناسب با اولویت ها و سلائق مشتری اقدام نمایند. این نوع نرم افزارها را سیستم های توصیه گر یا پیشنهاد دهنده می نامند[۳] [۶][۷].

از بین روش ها و الگوریتم های گوناگون ارائه پیشنهاد توسط سیستم های توصیه گر، معروف ترین و پر کاربردترین آنها روش فیلتر اشتراکی یا فیلتر مشارکتی[۴] می باشد. در این روش از ماتریس امتیاز کاربران به آیتم ها استفاده می شود. البته این روش علی رغم مزایای متعدد و کاربردهای فراوان، دارای نقاط ضعف هایی مانند شروع سرد[۵] و ماتریس خالی[۶] نیز می باشد [۶][۸].

روش پایه فیلتر محتوا محور[۷] نیز یکی از روش های مرسوم در این زمینه می باشد. این روش مبتنی بر استفاده از ویژگی های محصولات یا آیتم ها و شباهت بین آنها می باشد. این روش نسبت به روش فیلتر اشتراکی دارای سرعت عمل بالاتری می باشد ولی از نظر دقت پیش بینی امتیازات، دارای دقت پائین تری می باشد و از طرفی میزان دقت پیشنهادات آن بستگی به تعداد ویژگی هایی دارد که برای بدست آوردن همبستگی بین آیتم ها بکار برده می شود.

گستردگی محصولات رسانه ای از نظر تنوع و مخاطب، و از طرفی سرعت و سادگی تبادل این محصولات بین فروشنده و خریدار، باعث گردیده تا سیستم های توصیه گر در صنعت رسانه دارای کاربردهای فراوانی باشد [۱۰]. در بین محصولات رسانه ای نیز فیلم از جایگاه بالاتری نسبت به سایر محصولات برخوردار است. یکی از اتفاقات تاثیر گذار در گسترش روش ها و الگوریتم های سیستم های توصیه گر خصوصا" در زمینه فیلم، مسابقه یک میلیون دلاری شرکت نتفلیکس[۸] در اکتبر ۲۰۰۶، جهت ارتقاء سیستم توصیه گر فیلم خود بود. شرکت نتفلیکس اقدام به برگزاری مسابقه ای با هدف ارتقاء ۱۰٪ دقت پیشنهادات با روش فیلتر اشتراکی نمود. این مسابقه باعث گردید روش های نوظهور بسیاری برای سیستم های توصیه گر بر پایه روش فیلتر اشتراکی توسعه داده شود. جایزه یک میلیون دلاری این مسابقه پس از سه سال در سپتامبر ۲۰۰۹ به یک تیم تحقیقاتی که به ارتقاء دقت به میزان ۱۰,۰۶٪ دست یافته بودند، تعلق گرفت[۹] [۱۱].

سابقه تحقیقات دانشگاهی در موضوع سیستم های توصیه گر در دنیا، به حدود سه دهه قبل بر می گردد، ولی هنوز موضوع مذکور به جهت کاربردها و مزایای فراوان به ویژه در تجارت B2C، یکی از موضوعات تحقیقاتی به روز می باشد.

بررسی علمداری و همکاران (۲۰۲۰) [۲]، بر روی مقالات منتشر شده توسط جستجوگر گوگل اسکولار با جستجوی کلمات کلیدی "Recommendation Systems" و "E-Commerce" نشان می دهد از بین ۳۵۸ مقاله یافت شده در این موضوعات، ۲۸۹ مورد آن (بیش از ۸۰٪) بین سال های ۲۰۰۸ تا ۲۰۱۹ انتشار یافته اند که نشانگر فعال بودن موضوع تحقیق می باشد.

البته برخلاف چنین سابقه تحقیقاتی در دنیا، بررسی نویسندگان نشان می دهد تعداد کارهای تحقیقاتی در زمینه سیستم های توصیه گر در داخل کشور محدود بوده و در زمینه مقالات منتشر شده در مجلات معتبر داخلی برای سیستم های توصیه گر فیلم ، نتایج جستجو در گوگل اسکولار[۱۰]، ایران داک[۱۱] و مگیران[۱۲] و با کلمات کلیدی "*سیستم های توصیه گر فیلم* " و "*سیستم های پیشنهاد دهنده فیلم*"، منجر به یافتن تعداد بسیار اندکی مقاله (۵ مقاله) به زبان فارسی گردید. البته سیستم های توصیه گر در شرکت های معتبر داخلی مانند دیجی کالا، نماوا و فیلیمو مورد استفاده قرار می گیرد ولی در خصوص روش های بکار رفته و میزان دقت و کارائی آنها مطالب اندکی می توان یافت. در این خصوص، به عنوان یکی از کارهای ارزنده و به روز می توان به سیستم توصیه گر شرکت فیلیمو اشاره کرد که حاصل کار تیمی قوی می باشد [۱۲]، ولی در مقایسه با سایت های مشابه خارجی دارای ضعف های مشهودی می باشد.

هدف از ارائه این مقاله، نحوه پیاده سازی سیستم های توصیه گر فیلم، ارتقاء سیستم های مرسوم با روش های کم هزینه و برای سایت های کوچک و متوسط و قابل کاربرد در داخل کشور است. بنابراین، در این مقاله به ارائه مدلی از سیستم توصیه گر فیلم بر مبنای فیلتر اشتراکی و با استفاده از تکنیک های متنوع جهت ارتقاء دقت پیشنهادات پرداخته شده است. برای داده های ورودی از سایت معتبر گروپ لنز و از دیتاست مووی لنز استفاده شده است.

در ادامه مقاله، به مرور ادبیات ، پیشینه تحقیق، ، روش تحقیق و بررسی نتایج حاصل از مدل پیشنهادی پرداخته شده است.

---

۳ Recommender/Recommendation Systems

٤ Collaborative Filtering

٥ Cold Start

٦ Sparsity

۷ Content Based

۸ Netflix

۹ Netflixpriz.com

۱۰ Google Scholar

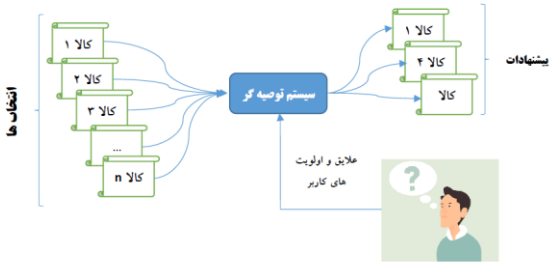
۱۱ Irandoc

۱۲ Magiran

۲. مرور ادبیات و پیشینه تحقیق

۱,۲ سیستم های توصیه گر

شکل ۱ سیستم های توصیه گر را بصورت شماتیک نمایش می دهد. این سیستم از بین انبوه انتخاب ها بهترین موارد را انتخاب و به کاربر پیشنهاد می نماید.



شکل ۱. سیستم توصیه گر

سیستم های توصیه گر به جهت استفاده از داده های کلان، روش های مختلف داده کاوی و لزوم استفاده از فیلتر کردن داده ها، در مجموع به سه نوع سیستم توصیه گر اصلی تقسیم می گردند: فیلتر محتوا- محور، فیلتر اشتراکی و فیلتر ترکیبی[۲][۶][۱۳].

سیستم های فیلتر اشتراکی نیز به دو نوع سیستم حافظه محور و مدل محور تقسیم می گردد. در سیستم حافظه محور از داده های موجود استفاده و با استفاده از شباهت بین کاربران وآیتم ها موارد پیشنهادی استخراج می گردد در حالیکه در روش مدل محور، از داده های موجود جهت یادگیری یک مدل پیشنهادی استفاده می گردد تا بتوان مقادیر نامعلوم را پیش بینی نمود.

از طرفی برای ارائه پیشنهادات به عنوان خروجی سیستم های توصیه گر به دو روش می توان اقدام نمود. روش استفاده از تشابه یا فاصله بین آیتم ها و یا کاربران و روش پیش بینی امتیازات کاربران به آیتم هایی که تاکنون استفاده نکرده است. در روش پیش بینی امتیازات، دقت نتایج بسیار مهم می باشد [۲۵].

۲,۲ اهداف و چالش های سیستم های توصیه گر

نویسندگان در [۶] معتقد است اصلی ترین هدف از بکارگیری سیستم های توصیه گر، افزایش فروش محصولات می باشد و این هدف با ارائه پیشنهادات مناسب به کاربران و جلب توجه آنها میسر می گردد. از طرفی برای رسیدن به هدف مذکور لازم است در طراحی سیستم های توصیه گر و ارائه پیشنهادات، جنبه ها و اهداف فنی مرتبط بودن[۱۳]، تازگی[۱۴]، موارد غیر منتظره[۱۵]و تنوع توصیه ها[۱۶] نیز رعایت گردند.

همچنین در طراحی سیستم های توصیه گر باید چالش های پیاده سازی هر یک از روش های سیستم های توصیه گر نیز مد نظر قرار گیرد. اهم چالش ها در سیستم های توصیه گر عبارتند از: استخراج محتوا، شروع سرد، عدم وجود اطلاعات، خالی بودن[۱۷] ماتریس امتیازات[۱۸] ، مقیاس پذیری[۱۹] ، حفظ حریم خصوصی و کلمات مترادف [۵][۶][۱۳]

نکته بسیار مهم در ارائه سیستم های توصیه گر عدم توانایی هیچ یک از روش ها، به برآورده نمودن همه اهداف و برطرف نمودن همه چالش های مذکور می باشد و هر کدام از روش ها در برآورده کردن اهداف و برطرف نمودن چالش ها، دارای نقاط قوت و ضعف می باشند. لذا برای انتخاب روش مناسب لازم است مشخصات آیتم ها، کاربران و اطلاعات موجود از آنها در نظر گرفته شود و با بررسی روش های مختلف نسبت به انتخاب روش بهینه با بیشترین دقت پیشنهادات و کمترین هزینه از نظر میزان حجم عملیات و حجم حافظه مورد نیاز، اقدام گردد [۲][۶].

۳,۲ پیشینه تحقیق

سیستم های توصیه گر فیلم از پر کاربردترین موارد استفاده از سیستم های توصیه گر می باشد و در این خصوص تحقیقات فراوانی انجام شده است. در ادامه به برخی موارد اشاره می گردد.

نویسندگان در [۱۴]، با استفاده از روش فیلتر اشتراکی و با کمک یادگیری عمیق[۲۰]، نسبت به ارائه مدلی برای رفع مشکل شروع سرد آیتم ها اقدام کرده اند. بررسی ها بر روی دیتاست نتفلیکس نشان از موثر بودن بکارگیری روش های فیلتر اشتراکی و یادگیری عمیق با یکدیگر برای ارائه پیشنهاد برای شروع سرد می باشد.

نویسندگان در [۱۵]، در تحقیق خود با ترکیب روش محتوا محور و فیلتر اشتراکی به صورت تقویتی، سیستم توصیه گری طراحی کرده اند که علاوه بر حل مشکل شروع سرد، مسئله اعتماد را نیز پوشش می دهد.

نویسندگان در [۱۶] از شبکه های عصبی برای استخراج ویژگی ها از داده های تصویری و عددی به عنوان ورودی به سایر مدل های یادگیری ماشین، مانند ماشین های بردار پشتیبان و نزدیکترین

---

۱۳ Relevance
۱۴ Novelty
۱۵ Serendipity
۱۶ Diversity
۱۷ Sparsity
۱۸ Rating Matrix
۱۹ Scalability
20 Deep Learning

همسایه استفاده کرده اند تا اثر ویژگی های استخراج شده بر روی مدل های مذکور بررسی گردد.

نویسندگان در [۱۷] در بررسی خود به ارائه سیستم توصیه گر محتوا محور با استفاده از همبستگی ژانر[۲۱] پرداخته اند. آنها در این روش با توجه به سابقه رفتار کاربر نسبت به فیلم ها، ژانرهای مورد علاقه شخص را تشخیص داده و فیلم های در ژانر مورد نظر را به او پیشنهاد داده اند.

نویسندگان در [۸] روش خوشه بندی k-means با الگوریتم K نزدیکترین همسایه[۲۲] را بر روی دیتاست فیلم ها اعمال کرده اند، تا کاربران را بر اساس امتیازات داده شده به فیلم های مختلف، خوشه بندی نموده و متوسط امتیاز K کاربر درون یک خوشه به هر فیلم را به عنوان امتیاز آن خوشه در نظر بگیرند. به هرا و همکاران[۲۳] (۲۰۱۹) [۱۸] با کمک روش شبکه عصبی ماشین بولتزمن محدود[۲۴] و استفاده از فیلتر اشتراکی، نسبت به تخمین مقادیر امتیازات ناموجود اقدام نموده و نتایج را با روش همبستگی پیرسون و متوسط وزندار مقایسه کرده اند.

نویسندگان در [۱۹] از دو روش فیلتر اشتراکی و شبکه های عصبی و از کتابخانه های زبان برنامه نویسی پایتون[۲۵] استفاده کرده و عملکرد آنها را از نظر سرعت و دقت با هم مقایسه کرده اند. در این تحقیق دو سیستم در نظر گرفته شده است. در اولین سیستم، از فیلتر اشتراکی مدل محور و روش KNN و K-Means برای پیش بینی امتیازات کاربران استفاده شده است. در سیستم دوم، از بسته های پایتون sk_learn و tensorflow استفاده شده است.

نویسندگان در [۲۰] از روش ترکیبی محتوا محور و فیلتر اشتراکی استفاده کرده اند تا از محدودیت های ناشی از روش محتوا محور رهایی یابند. در روش پیشنهادی، آنها از یک پیش کانال قبل از اعمال الگوریتم K-Mean استفاده کرده اند. برای تفکیک فاصله هر نقطه از مرکز، از ژانر و امتیاز استفاده شده است. در این روش فقط برای کاربرانی که حداقل به شش فیلم امتیاز داده اند، سیستم توصیه گر پیشنهاد ارائه می دهد. پیشنهادات از طریق روش فیلتر اشتراکی محاسبه و ارائه می شود.

نویسندگان در [۲۱] در تحقیق خود از روشی تحت عنوان MCBF-SVD را برای پیش بینی امتیازات کاربران فعال، به فیلم ها بر مبنای داده های صریح[۲۶] و ضمنی[۲۷] ناشی از دیتاست فیلم ها ارائه می دهند. روش فوق از عوامل وزن دار برای بررسی اثر انواع دسته بندی فیلم بر روی پیش بینی امتیازات کاربر و همچنین ارتقاء روش فیلترینگ بر مبنای دسته بندی فیلم ها استفاده می کند. در نهایت، روش MCBF با الگوریتم تجزیه ماتریس به مقادیر منفرد[۲۸] ترکیب می شود تا از عملکرد خوب آن در فیلتر اشتراکی بتوان استفاده کرد. در مقایسه با انواع الگوریتم های مختلف موجود، روش پیشنهادی MCBF-SVD به میزان قابل توجهی دقت پیش بینی امتیازدهی را ارتقاء می بخشد و مقیاس پذیری و اثربخشی سیستم های توصیه گر شخصی را افزایش می دهد.

بررسی پیشینه تحقیق نشان می دهد توسعه تجارت الکترونیک مستلزم استفاده از سیستم های توصیه گر می باشد [۱۴][۲۲] و روش های بکارگیری سیستم توصیه گر نیز به حوزه کاری وابسته می باشد و هر گروه محصول یا خدمات نیازمند سیستم توصیه گر مخصوص به خود می باشند، زیرا ویژگی های محصولات با یکدیگر متفاوت می باشند [۲].

از طرفی سیستم های توصیه گر با حجم عظیمی از داده ها (کلان داده) که در حال افزایش نیز می باشد سر و کار دارند و باید قادر به انجام محاسبات با هزینه معقول باشند [۷][۲۳].

تاکید می گردد هیچ یک از سیستم های توصیه گر قادر به حل تمام چالش ها نمی باشد و بسته به موضوع و ویژگی های کاربران و آیتم ها، لازم است روش های مختلف ارزیابی شده و بهترین روش انتخاب گردد [۲]. در این بین، سیستم توصیه گر فیلتر اشتراکی با توجه به تنوع الگوریتم ها و دقت مناسب، دارای بیشترین کاربرد می باشد و در این مقاله نیز مورد استفاده قرار خواهد گرفت. البته در اغلب مواقع لازم است این روش با سایر روش ها ترکیب گردد تا بر مشکلاتی مانند شروع سرد و خالی بودن ماتریس امتیازات غالب گردد [۹][۱۴][۲۴].

۳. روش تحقیق

هدف از این تحقیق ارائه مدلی از سیستم توصیه گر فیلم بر اساس فیلتر اشتراکی و با استفاده از روش محاسبه ضریب همبستگی پیرسون بین کاربران می باشد تا هم بر چالش خالی بودن ماتریس امتیازات غلبه نماید و هم دارای خطای قابل قبولی در مقایسه با سایر روش ها باشد. برای اینکار از سه روش برای پر کردن جاهای خالی، پیدا کردن بهترین تعداد نزدیکترین همسایه ها و محاسبه پیش بینی امتیازات بر اساس میانگین وزنی استفاده شده و در نهایت روش بهینه انتخاب شده است.

21 Genre Correlation
22 K Nearest Neighbors
23 Behera, D.K., et al.
24 Restricted Boltzmann Machine
25 Python
26 Explicit
27 Implicit
28 Singular Value Decomposition

فرایند کار به شرح ذیل می باشد:

۱- حذف فیلم های با تعداد امتیاز ۱۰ و یا کمتر

۲- تقسیم داده های موجود به دو مجموعه داده آموزش و تست به نسبت ۸۰ به ۲۰ بصورت تصادفی

۳- پر کردن جاهای خالی با مقادیر زیر:

الف: میانگین امتیاز هر کاربر ب: میانگین امتیا هر فیلم

ج: میانگین امتیازکل

۴- محاسبه ضریب همبستگی بر اساس روش پیرسون

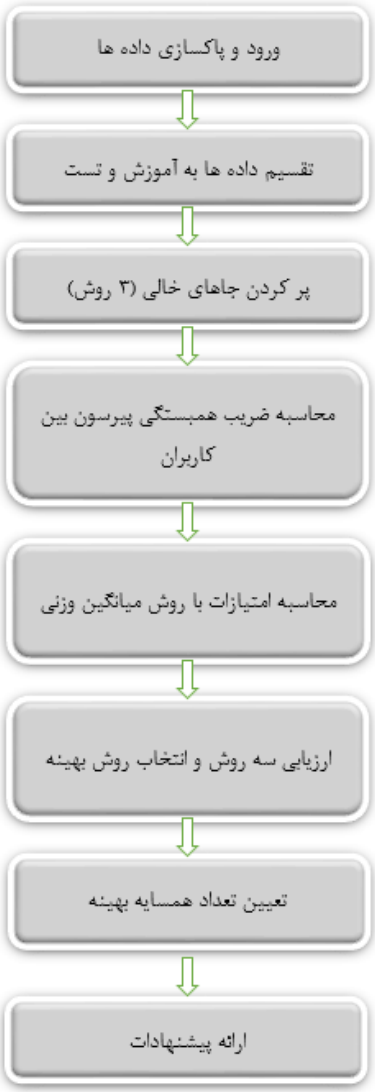
۵- محاسبه پیش بینی امتیازات با روش میانگین وزنی

۶- ارزیابی سه روش و انتخاب بهترین روش

۷- تعیین تعداد همسایه های بهینه

۸- ارائه پیشنهادات

شکل ۲ نمودار شماتیک مراحل انجام کار را نشان می دهد:



شکل ۲. فرایند روش تحق

۴. یافته های تحقیق

۱,۴ جمع آوری داده ها

برای داده های ورودی از مجموعه داده ml-latest-small مووی لنز[۲۹] با ۶۱۰ کاربر ، ۹۷۴۲ فیلم و حدود ۱۰۰ هزار امتیاز استفاده شده است. این داده ها متعلق به فاصله زمانی مارچ ۱۹۹۶ تا سپتامبر ۲۰۱۸ می باشند. تعدادی از فیلم ها به جهت عدم اشتراک در امتیازدهی و یا با تعداد ۱۰ امتیاز یا کمتر از مجموعه داده ها حذف گردید. مجموعه داده های اولیه در شکل ۳ نشان داده شده است:

۲۹ http://grouplens.org/datasets/

	userId	movieId	rating
0	1	1	4.0
1	1	3	4.0
2	1	6	4.0
3	1	47	5.0
4	1	50	5.0
...	...	...	...
100831	610	166534	4.0
100832	610	168248	5.0
100833	610	168250	5.0
100834	610	168252	5.0
100835	610	170875	3.0

100836 rows × 3 columns

شکل ۳. داده های اولیه

'test'

	movieId	userId	rating
78754	2918	610	3.5
33967	94959	280	5.0
11513	7451	89	2.5
28751	1587	234	3.0
63044	64839	490	4.0
...	...	...	...
19715	3578	166	3.5
58650	14	474	3.0
43014	3910	346	3.5
1324	750	16	4.5
48129	1013	385	3.0

15869 rows × 3 columns

شکل ۴. داده های آموزش و تست

۲,۴ پاکسازی داده ها

با توجه به اینکه داده ها فاقد کاربران با تعداد امتیاز کمتر از ۲۰ می باشند، لذا فقط کاربر با شناسه ۵۳ که تمام امتیازات آن یکسان می باشد و باعث عدم محاسبه ضریب همبستگی می گردد و همچنین فیلم های دارای تعداد امتیاز مساوی و کمتر از ۱۰ حذف شده اند. این عمل باعث می گردد تا کاربران و فیلم های دارای تاثیر کمتر حذف شده و از طرفی حجم حافظه و محاسبات نیز کاهش یابد.

۳,۴ محاسبات و نتایج تحقیق

در این مرحله، امتیازات بصورت تصادفی و به نسبت ۸۰ به ۲۰ بین داده های آموزش و تست تقسیم می گردد. شکل ۴ این داده ها را نمایش می دهد:

'train'

	movieId	userId	rating
42589	88125	339	2.5
19434	380	161	3.5
57701	227	464	4.0
43403	3113	352	4.0
8919	1370	68	2.5
...	...	...	...
63370	3977	495	2.0
65615	3911	522	2.5
77655	46965	606	1.0
56088	6059	448	3.5
38408	1639	307	4.5

63475 rows × 3 columns

سپس میانگین امتیازات هر کاربر، هر فیلم و میانگین کل امتیازات بصورت جداگانه محاسبه شده و در سه مرحله در جاهای خالی ماتریس امتیازات حاصل از داده های آموزش قرار می گیرد تا محاسبه میانگین وزنی امتیازات و خطاها صورت پذیرد. نمونه حاصل از قرار دادن میانگین کاربر در شکل ۵ نمایش داده شده است:

movieId	1	2	3	5	6	7
userId						
1	4.000000	4.404908	4.000000	4.404908	4.000000	4.404908
2	4.022727	4.022727	4.022727	4.022727	4.022727	4.022727
3	1.611111	1.611111	1.611111	1.611111	1.611111	1.611111
4	3.479730	3.479730	3.479730	3.479730	3.479730	3.479730
5	4.000000	3.705882	3.705882	3.705882	3.705882	3.705882
...	...	...	...	...	...	...
606	2.500000	3.689189	3.689189	3.689189	3.689189	2.500000
607	4.000000	3.741007	3.741007	3.741007	3.741007	3.741007
608	2.500000	2.000000	2.000000	3.223729	3.223729	3.223729
609	3.000000	3.240000	3.240000	3.240000	3.240000	3.240000
610	5.000000	3.811419	3.811419	3.811419	5.000000	3.811419

596 rows × 2119 columns

شکل ۵. ماتریس امتیازات پس از جاگذاری میانگین امتیازات کاربر

برای محاسبه میانگین وزنی امتیازات، لازم است تا ضریب همبستگی پیرسون بین کاربران محاسبه گردد. لذا با استفده از رابطه ۱ ماتریس همبستگی شکل ۶ بدست می آید.

$$S_{Pearson}(u,v)=\frac{\sum_{i=1}^{n}(r_{u,i}-\bar{r}_u)\times(r_{v,i}-r_v)}{\sqrt{\sum_{i=1}^{n}(r_{u,i}-\bar{r}_u)^2\times(r_{v,i}-\bar{r}_v)^2}} \qquad (۱)$$

userId	1	2	3	4	5	6	7	8
userId								
1	1.000	-0.001	0.025	0.067	0.009	-0.080	-0.007	0.055
2	-0.001	1.000	-0.000	-0.000	-0.000	-0.035	0.001	-0.068
3	0.025	-0.000	1.000	0.000	0.000	0.009	0.000	-0.035
4	0.067	-0.000	0.000	1.000	-0.059	0.016	0.067	0.022
5	0.009	-0.000	0.000	-0.059	1.000	0.049	0.007	-0.015
...	...	...	...	...	...	...	...	...
606	0.039	0.000	-0.086	0.022	0.021	-0.008	0.039	0.026
607	0.055	-0.033	-0.011	0.003	0.020	0.057	0.041	0.052
608	0.067	-0.016	-0.013	-0.056	-0.005	0.009	0.023	0.050
609	-0.062	-0.106	-0.000	-0.021	0.128	0.085	0.032	0.059
610	0.018	0.035	-0.000	-0.007	0.008	0.020	0.057	0.033

596 rows × 596 columns

شکل ۶. بخشی از ماتریس همبستگی کاربران

در این مرحله، با کمک ماتریس همبستگی و ماتریس امتیازات و با استفاده از رابطه ۲، میانگین وزنی امتیازات را بر تغییر تعداد همسایه های هر کاربر از یک تا ده ، محاسبه می کنیم.

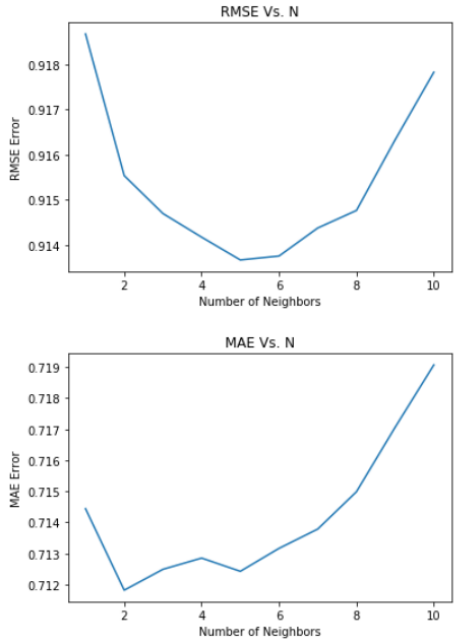
$$W = \frac{\sum_{i=1}^{n}(w_i * r_i)}{\sum_{i=1}^{n} w_i} \qquad (۲)$$

در مرحله بعد، با استفاده از روش های ریشه میانگین مربعات خطا (RMSE) و میانگین قدر مطلق خطا (MAE)، مقادیر پیش بینی شده و مقادیر داده های تست را ارزیابی می کند. برای محاسبه این مقادیر از روابط ۳ و ۴ استفاده شده است:

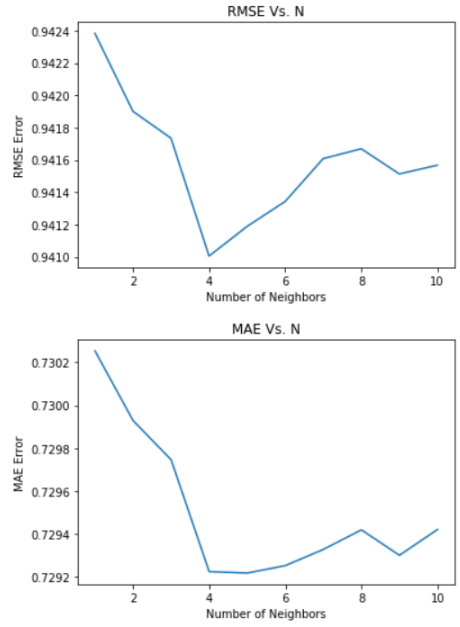
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(r_i - p_i)^2}{n}} \qquad (۳)$$

$$MAE = \frac{\sum_{i=1}^{n}|r_i - p_i|}{n} \qquad (۴)$$

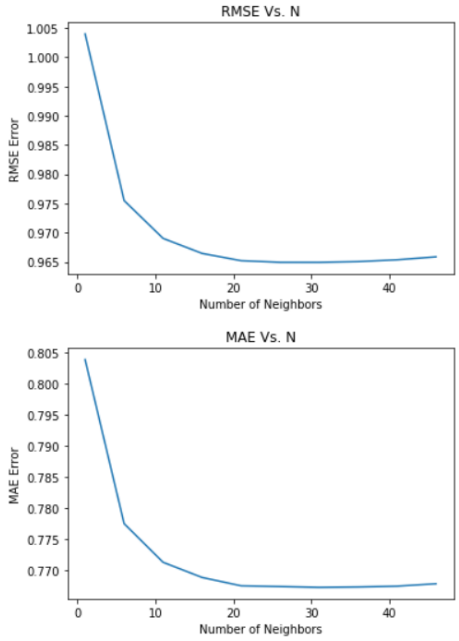
در این مرحله با توجه به مقادیر بدست آمده برای خطای RMSE بهترین روش انتخاب می گردد. همانطور که از شکل های ۷ تا ۹ مشخص است، روش قرار دادن میانگین امتیاز هر کاربر دارای کمترین خطا در بین سه روش می باشد.



شکل ۷. نمودارهای خطا در حالت قرار گرفتن متوسط امتیاز کاربر بجای مقادیر خالی



شکل ۸. نمودارهای خطا در حالت قرار گرفتن متوسط امتیاز فیلم بجای مقادیر خالی



شکل ۹. نمودارهای خطا در حالت قرار گرفتن متوسط امتیاز کل بجای مقادیر خالی

حال با توجه به روش انتخاب شده (میانگین امتیاز کاربر)، تعداد همسایه های بهینه با توجه به منحنی RMSE تعیین می گردد. همانطور که از شکل ۷ مشخص است بهترین تعداد همسایه ۵ همسایه می باشد که کمترین مقدار خطای پیش بینی آن برابر با ۰,۹۱۳۶۷ می باشد. شکل ۱۰ مقادیر خطا به ازای تعداد مختلف همسایه ها را نمایش می دهد.

	MAE	RMSE	neighbs_no
0	0.714442	0.918669	1.0
1	0.711818	0.915533	2.0
2	0.712490	0.914694	3.0
3	0.712851	0.914171	4.0
4	0.712425	0.913666	5.0
5	0.713165	0.913753	6.0
6	0.713786	0.914374	7.0
7	0.714988	0.914762	8.0
8	0.717064	0.916326	9.0
9	0.719072	0.917822	10.0

شکل ۱۰. تغییر میزان خطا بر اثر تعداد مختلف همسایگی

شکل ۱۱ بخشی از جدول امتیازات پیش بینی شده با تعداد ۵ همسایه را نمایش می دهد:

movieId	1	2	3	5	6	7
userId						
1	4.000000	4.061152	4.000000	4.041837	4.000000	4.041837
2	3.980667	3.946274	3.946274	3.946274	3.946274	3.946274
3	2.282680	2.282680	2.282680	2.282680	2.245672	2.282680
4	3.372163	3.487317	3.487317	3.487317	3.487317	3.487317
5	4.000000	3.526633	3.489042	3.569444	3.569444	3.569444
...	...	...	...	...	...	...
606	2.500000	3.528934	3.628002	3.537123	3.826501	2.500000
607	4.000000	3.590499	3.654838	3.654838	3.654838	3.654838
608	2.500000	2.000000	2.000000	3.152897	3.385019	3.258558
609	3.000000	3.371780	3.371780	3.371780	3.401956	3.371780
610	5.000000	3.697487	3.606409	3.581372	5.000000	3.570124

596 rows × 2119 columns

شکل ۱۱. ماتریس امتیازات پیش بینی شده

در مرحله آخر، به عنوان خروجی سیستم توصیه گر پیشنهادی، برای کاربر با شناسه ۲۵۰، اقدام به ارائه ۱۰ فیلم پیشنهادی با بالاترین امتیاز پیش بینی شده می نماییم که خروجی آن بصورت شکل ۱۲ می باشد:

movieId	250	title
593	4.273479	Silence of the Lambs, The (1991)
457	4.273479	Fugitive, The (1993)
356	4.246928	Forrest Gump (1994)
296	4.226642	Pulp Fiction (1994)
32	4.226642	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
1	4.179805	Toy Story (1995)
34	4.179805	Babe (1995)
47	4.179805	Seven (a.k.a. Se7en) (1995)
4993	4.175357	Lord of the Rings: The Fellowship of the Ring,...
1222	4.159424	Full Metal Jacket (1987)

شکل ۱۲. ده فیلم پیشنهادی به کاربر با شناسه ۲۵۰

۵. بحث و نتیجه گیری

۱,۵ بحث و بررسی

بهترین نتایج حاصل از سه نوع جاگذاری انجام شده در جدول (۱) نمایش داده شده است:

جدول ۱. بهترین مقادیر خطا در سه روش جاگذاری

روش	RMSE	MAE	N
میانگین کاربر	۰/۹۱۳۷	۰/۷۱۲۴	۵
میانگین فیلم	۰/۹۴۱۰	۰/۷۲۹۲	۴
میانگین کل	۰/۹۶۴۹	۰/۷۶۷۳	۳۱

بهترین نتیجه هنگامی رخ می دهد که بجای مقادیر نامعلوم، میانگین امتیازات هر کاربر در هر سطر قرار گیرد.

بنابراین میزان خطای RMSE در روش پیشنهادی، برابر با ۰/۹۱۳۷ می باشد که در مقایسه با بهترین نتیجه بدست آمده از مقاله ویدیانینگ تیاس (۲۰۲۱) [۲۷] که برای RMSE مقدار ۰/۹۴۴۸ را بدست آورده است، ۳/۲۹٪ خطای کمتری را دارا می باشد. لازم به ذکر است مقاله مذکور از دیتاست ml-۱۰۰k مووی لنز با ۹۳/۷ ٪ خالی بودن ماتریس و با تعداد ۹۴۳ کاربر و ۱۶۸۲ فیلم استفاده کرده است در حالیکه نسخه مورد استفاده در تحقیق حاضر دارای ۹۸/۳٪ ماتریس خالی می باشد.

استفاده از ضریب همبستگی در تولید سیستم های توصیه گر یکی از روش های متداول و موثر می باشد و علی رغم انواع روش های مختلف بکار گرفته شده در این خصوص تاکنون، ولی همچنان جزء روش های محبوب می باشد [۲۷]. این روش با توجه به امکان استفاده از ضریب همبستگی بین کاربران و آیتم ها و همچنین امکان ادغام الگوریتم های مختلف از جمله KNN و K_Means و از طرفی امکان قرار دادن مقادیر مختلف در جاهای خالی، امکان تولید انواع مختلفی از سیستم های توصیه گر فیلتر اشتراکی را فراهم می کند.

۲,۵ جمع بندی و پیشنهادات

تحقیق حاضر نشان می دهد پیاده سازی سیستم توصیه گر فیلتر اشتراکی بر مبنای ضریب همبستگی در صورت ادغام با روش هایی مانند KNN و K_Means و بکار بردن ابتکارات لازم قادر به ارائه پیشنهادات شخصی سازی شده مناسب و با دقت کافی می باشد.

پیشنهاد می گردد در ادامه کار مشکل شروع سرد در این روش مورد بررسی قرار گیرد. همچنین ترکیب الگوریتم های مختلف شبکه عصبی مانند شبکه های چند لایه پرسپترون (MLP) ، ماشین بولتزمن محدود (RBM) و الگوریتم تجزیه ماتریس به مقادیر منفرد (SVD) با روش مذکور مورد بررسی قرار گیرد.

مراجع

[1] H. Øverby and Jan A. Audestad, 2018, Digital Economics. .

[2] P. M. Alamdari, N. J. Navimipour, M. Hosseinzadeh, A. A. Safaei, and A. Darwesh, "A Systematic Study on the Recommender Systems in the E-Commerce," IEEE Access, vol. 8, pp. 115694–115716, 2020, doi: 10.1109/ACCESS.2020.3002803.

[3] Y. Zhang, H. Abbas, and Y. Sun, "Smart e-commerce integration with recommender systems," Electronic Markets, vol. 29, no. 2. pp. 219–220, 2019, doi: 10.1007/s12525-019-00346-x.

[4] G. Lekakos and P. Caravelas, "A hybrid approach for movie recommendation," Multimed. Tools Appl., vol. 36, no. 1–2, pp. 55–70, 2008, doi: 10.1007/s11042-006-0082-7.

[5] S. K. Raghuwanshi and R. K. Pateriya, "Recommendation systems: Techniques, challenges, application, and evaluation," in Advances in Intelligent Systems and Computing, vol. 817, 2019, pp. 151–164.

[6] Aggarwal, C.C., Recommender systems. Vol. 1. 2016: Springer

[7] R. Singla, S. Gupta, A. Gupta, and D. K. Vishwakarma, "FLEX: A content based movie recommender," 2020 Int. Conf. Emerg. Technol. INCET 2020, pp. 8–11, 2020, doi: 10.1109/INCET49848.2020.9154163.

[8] R. Ahuja, A. Solanki, and A. Nayyar, "Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor(2019).pdf," Proc. 9th Int. Conf. Cloud Comput. Data Sci. Eng. Conflu. 2019, pp. 263–268, 2019, doi: 10.1109/CONFLUENCE.2019.8776969.

[9] G. Geetha, M. Safa, C. Fancy, and D. Saranya, "A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System," J. *Phys. Conf. Ser.*, vol. 1000, no. 1, 2018, doi: 10.1088/1742-6596/1000/1/012101.

[10] S. M. Choi, S. K. Ko, and Y. S. Han, "A movie recommendation algorithm based on genre correlations," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 8079–8085, 2012, doi: 10.1016/j.eswa.2012.01.132.

[11] en.wikipedia.org/wiki/Netflix_prize

[12] https://aparat.design/recommendation-systems-in-filimo-cqzge4gfsevi

[13] S. M. Choi, S. K. Ko, and Y. S. Han, "A movie recommendation algorithm based on genre correlations," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 8079–8085, 2012, doi: 10.1016/j.eswa.2012.01.132.

[14] Z. T. Jian Wei a , Jianhua He a , ∗, Kai Chen b , Yi Zhou c, "Collaborative filtering and deep learning based recommendation system for cold start items." 2016.

[۱۵]حیدری, ب., پروین نیا,ا., ۱۳۹۶، ارائه مدلی برای سیستم های توصیه گر فیلم مبتنی بر رویکرد مشارکت محور، مجله فناوری اطلاعات در طراحی مهندسی، دوره ۱۰، شماره ۱، شهریور ۱۳۹۶، صفحه ۱ تا ۹

[16] Notley, S. and M. Magdon-Ismail, Examining the use of neural networks for feature extraction: A comparative analysis using deep learning, support vector machines, and k-nearest neighbor classifiers. arXiv preprint arXiv:1805.02294, 2018

[17] S. Reddy, S. Nalluri, S. Kunisetti, S. Ashok, and B. Venkatesh, "Content-based movie recommendation system using genre correlation," in *Smart Innovation, Systems and Technologies*, 2019, vol. 105, pp. 391–397, doi: 10.1007/978-981-13-1927-3_42.

[18] Behera, D.K., M. Das, and S. Swetanisha, Predicting Users' Preferences for Movie Recommender System Using Restricted Boltzmann Machine, in Computational Intelligence in Data Mining. 2019, Springer. p. 759-769

[19] C. H. Lin and H. Chi, A Novel Movie Recommendation System Based on Collaborative Filtering and Neural *Networks*, vol. 926. Springer International Publishing, 2020.

[20] Furtado, F. and A. Singh, Movie recommendation system using machine learning. International Journal of Research in Industrial Engineering, 2020. 9(1): p. 84-98

[21] X. Li, H. Zhao, Z. Wang, and Z. Yu, "Research on Movie Rating Prediction Algorithms," 2020 5th IEEE Int. Conf. Big Data Anal. ICBDA 2020, pp. 121–125, 2020, doi: 10.1109/ICBDA49040.2020.9101282.

[22] D. Lee and K. Hosanagar, "Impact of recommender systems on sales volume and diversity," 35th Int. Conf. Inf. Syst. "Building a Better World Through Inf. Syst. ICIS 2014, pp. 1–15, 2014.

[23] F. Ricci, L. Rokach, B. Shapira, P. B. Kantor, and F. Ricci, Recommender Systems Handbook. 2011.

[24] K. Falk, "Practical Recommender Systems." p. 406, 2019.

[۲۵]حسینی، م.؛ نصرالهی، م.؛ بقایی، ع.۱۳۹۷، یک سامانه توصیه گر ترکیبی با استفاده از اعتماد و خوشه بندی دو جهته به منظور افزایش کارائی پالایش گروهی

[۲۶] صابری، ن.؛ منتظر، غ. ۱۳۸۹، شخصی سازی محیط یادگیری الکترونیکی به کمک توصیه گر فازی مبتنی برتلفیق سبک یادگیری و سبک شناختی، فصلنامه علمی-پژوهشی فناوری اطلاعات و ارتباطات ایران، سال دوم، شماره های ۳و۴

[27] Widiyaningtyas, T., Hidayah, I., & Adji, T. B. (2021). User profile correlation-based similarity (UPCSim) algorithm in movie recommendation system. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00425-x