ترجمه فا

TarjomeFa.Com

ارائه شده توسط:

سایت ترجمه فا

مرجع جدیدترین مقالات ترجمه شده

از نشریات معتبر

یک تحقیق درباره سیستم های پیشنهاد دهنده مبتنی بر فیلترسازی براساس همکاری برای اپلیکیشن های اینترنتی موبایل

چکیده

با توسعه سریع و کاربرد اینترنت موبایل، مقدار عظیمی از داده های کاربر تولید شده و هر روزه جمع اوری می شود. چگونگی استفاده از مزایای کامل این اطلاعات در همه جا، در حال تبدیل شدن به جنبه اساسی سیستم پیشنهادی است. فیلترسازی مبتنی بر همکاری یا CF به طور وسیعی مطالعه شده و برای پیشگویی علایق کاربران موبایل و استفاده از پیشنهادات مناسب استفاده می شود. در این مقاله، ما ابتدا یک چارچوب سیستم پیشنهاددهنده CF را براساس داده های کاربر مختلف مطرح می کنیم که شامل درجه بندی های کاربر و رفتارهای کاربر می باشد. ویژگی های کلیدی این دو نوع داده ها مورد بحث قرار می گیرد. بعلاوه، چندین الگوریتم CF معمولی طبقه بندی می شود حین اینکه شیوه های مبتنی بر حافظه و شیوه های مبتنی بر مدل با هم مقایسه شود. دو مطالعه موردی در یک تلاش برای روایی سازی چارچوب مطرح شده ارائه می شود.

کلیدواژه ها: اینترنت موبایل، سیستم پیشنهاد دهنده، فیلترسازی مبتنی بر همکاری.

۱-مقدمه

همراه با توسعه سریع اینترنت موبایل و محاسبه ابری، مقادیر زیادی داده هر روز توسط افراد و ماشین ها تولید می شود. جامعه ما در واقع به عصر داده های کلان وارد می شود. به یمن وسایل هوشمند مختلف و اپلیکیشن های موبایل، کاربران اینترنت می توانند همه نوع اطلاعات را درباره اموزش، خرید، فعالیت اجتماعی و غیره بدست آورند. ولیکن حین اینکه حجم داده ها افزایش می یابد، افراد مجبورند با مسئله اطلاعات بیش از حد مواجه شوند، که انجام تصمیمات درست را سخت تر می سازد. این پدیده را به شکل افزایش بار اطلاعات می نامند. بعلاوه، کاربرانی که با توانایی ورودی وسایل موبایل محدود شده اند، معمولا نمی خواهند که تعداد زیادی لغت تایپ کنند تا بگویند که چه می خواهند. سیستم پیشنهاددهنده می تواند این مسائل را با تعریف موثر الزامات احتمالی کاربران و انتخاب گزینه های دلخواه از یک مقدار عظیمی از اطلاعات کاندیدا برطرف سازد. سیستم های پیشنهاد دهنده معمولا به دو طبقه طبقه بندی می شوند یعنی فیلترسازی مبتنی بر محتوا و مبتنی بر همکاری.

سیستم پیشنهاددهنده مبتنی بر محتوا از محتواهای گزینه ها استفاده کرده و شباهت ها را در میان آنها می یابند. بعد از تحلیل تعداد کافی گزینه هایی که یک کاربر قبلا آنرا دوست داشته، مشخصات علایق کاربر مشخص می شود. بعد سیستم پیشنهاددهنده می تواند پایگاه داده ها را جستجو کند و گزینه های مناسبی را طبق این مشخصات انتخاب کند. مشکل این الگوریتم ها در این امر نهفته است که چگونه ترجیحات کاربر را براساس محتویات گزینه ها بیابیم. بسیاری شیوه ها برای حل این مسئله در حیطه های داده کاوی و یادگیری ماشینی تدوین شده است. برای مثال، برای پیشنهاد برخی مقالات به خواننده ویژه ای، یک سیستم پیشنهاد دهنده ابتدا همه کتابهایی که این خواننده خوانده را بدست می اورد و بعد محتویات آنها را تحلیل می کند. کلمات کلیدی می تواند از متن با کمک روشهای متن کاوی استخراج شود مانند TF-IDF معروف. بعد از ملحق سازی توزین های مربوط به همه کلیدواژه ها به آنها، یک کتاب می تواند با یک بردار چندبعدی نمایش داده شود. الگوریتم های خوشه گیری ویژه می تواند اجرا شود تا این بردارها را بیابد که نمایانگر علایق این خواننده می باشد.

از سوی دیگر، فیلترسازی مبتنی بر همکاری یا CF یکی از تاثیرگزارترین الگوریتم های پیشنهاددهنده شده است. برخلاف شیوه های مبتنی بر محتوا، CF تنها متکی بر درجه بندی های گزینه از هر کاربر می باشد. این امر برپایه این فرضیه است که کاربرانی که همان گزینه ها را با درجه بندی های مشابهی درجه بندی می کنند احتمالا ترجیحات مشابهی دارند. CF به طور اخص طراحی شده تا پیشنهاداتی را فراهم سازد زمانی که اطلاعات مفصل درباره کاربران و گزینه ها غیرقابل دسترس باشد. بعلاوه، به طور موفقیت آمیزی مسئله تخصص سازی بیش از حد است که کاملا در سیستم های مبتنی بر محتوا متداول می باشد. تخصص سازی بیش از حد پدیده ای است که گزینه های توصیه شده همیشه همان بوده و تنوع پیشنهادات نادیده گرفته می شوند. حین اینکه CF پیشنهاداتی را طبق مجاورت ها انجام می دهند (افراد با ترجیحات مشابه)، گزینه ای که یک کاربر مصرف می کند می تواند تا حدی برای مجاورانش جدید باشد. ویژگی های بالا به ویژه جذابیت دارد که باعث می شود الگوریتم های CF تا حد زیادی در سیستم های پیشنهاد دهنده بکار گرفته شود.

اما، تا انجا که می دانیم، هر مطالعه معدودی ویژگی های متداول الگوریتم های مختلف CF را برای اپلیکیشن های اینترنت موبایل اشکار کرده اند. بعلاوه، اغلب تحقیقات موجود صرفا اصول الگوریتم های CF را معرفی کرده و اهمیت مطالعه موردی را نادیده گرفته که می تواند عملکردهای الگوریتم های معمولی را به طور بصری و به طور

ویژه نمایش دهد. از اینرو، این مقاله بر سیستم های پیشنهاد کننده مبتنی بر فیلترسازی مبتنی بر همکاری برای اپلیکیشن های اینترنت موبایل متمرکز است. بویژه، نقش های اصلی این مقاله به ترتیب ذیل روشن شده اند تا از داده های جمع اوری شده استفاده کنند و پیشنهادات مناسبی را ایجاد کنند. ویژگی های داده های جمع اوری شده از هر دو رفتارهای کاربر و درجه بندی های کاربر همچنین مورد بحث قرار گرفته و مقایسه شده اند.

-الگوریتم های CF طبقه بندی شده اند. روشهای اصلی CF به اختصار خلاصه سازی و معرفی شده اند.

-دو مطالعه موردی برای روایی سازی چارچوب مطرح شده ارائه شده اند. ارزیابی ها درباره الگوریتم های CF نماینده براساس پایگاه های داده دنیای واقعی با تحلیل و مقایسه مفصل اجرا شده اند.

بقیه این مقاله به ترتیب ذیل سازماندهی می شود. بخش دوم نمایانگر چارچوب CF می باشد. هر دو طبقه بندی و روشهای اصلی الگوریتم های CF معمولی در بخش سوم معرفی شده است. در بخش چهارم، ما دو مطالعه موردی را براساس پایگاه داده های دنیای واقعی برای تحلیل عملکردهای الگوریتم های CF اجرا کرده ایم. سرانجام اینکه، بخش پنجم قسمت نتیجه گیری این مقاله می باشد.

۲-چارچوب سیستم پیشنهاددهنده CF

طبق شکل ۱، چارچوب یک سیستم پیشنهاددهنده CF معمولی شامل اینهاست: ۱) گرداوری داده ها، ۲) پیش پردازش ۳) فیلترسازی مبتنی بر همکاری. اول اینکه، داده های کاربر از طریق سبکه های بی سیم جمع اوری شده و در پایگاه های داده ابری توده ای جمع اوری می شود. بعد برخی عملیات پیش پردازش برای تضمین یکپارچه سازی داده ها و قابلیت اتکا واجب است. براساس این داده ها، الگوریتم های CF برای پیشگویی علایق کاربر و پیشنهاد گزینه های مربوطه برای صرفه جویی در کار و زمان اجرا می شود.

A-گردآوری داده ها

گرداوری داده ها بنیان سیستم پیشنهادی کاملی می باشد. داده های جمع اوری شده اساسا به چهار دسته طبقه بندی می شود. داده های جمع آوری شده اساسا به چهار دسته طبقه بندی می شود: داده های فردی، داده های تولید، رفتار کاربر و درجه بندی کاربر.

۱)داده های فردی: بسیاری شرکتهای تجاری ملزم می کنند که کاربران در سرورهایشان ثبت نام کنند و اطلاعات فردی را قبل از استفاده از خدمات پر نمایند. اطلاعات فردی معمولا شامل نام، تلفن، جنسیت، سرگرمی ها و غیره می باشد. براساس تحلیل داده های فردی فوق، شرکتهای تجاری می توانند مشخصات کاربر را تعیین کرده و پیامهای تبلیغاتی را به مشتریان موبایل به طور اخص تر ارسال دارند.

۲) داده های تولید: تاجران تمایل به طبقه بندی کالاهای خود طبق عملکرد، برند، قیمت و غیره در آنها دارند. برای مثال، یک وب سایت ویدئویی معمولا برچسب هایی را به ویدئوهای خودش برای کمک به مصرف کنندگان جهت یافتن راحتتر از آنچه لذت می برند اضافه می نماید. با اینحساب، داده های تولید دسترسی راحتی توسط شرکتهای تجاری دارند.

۳) رفتار کاربر: حین مرور یک وب سایت یا گوش دادن به یک قطعه موسیقی، کاربران احتمالا تحت نظارت سرور هستند که مقدار زیادی داده های رفتاری را ذخیره سازی می کند از جمله طول گوش دادن به موسیقی، تاریخ خرید یک کتاب، یا حتی تعداد کلیک های یک وب سایت. این داده ها معمولا حجم زیادی داشته و باید با روشهای داده کاوی مخصوصی انالیز شوند.

۴) درجه بندی کاربر: برخی وب سایت ها سیستم های درجه بندی را ارائه داده اند و مصرف کنندگان را هدایت می کند تا گزینه هایی را درجه بندی کند که تجربه کرده اند مانند فیلم ها، آهنگ ها، و خدمات وب. این درجه بندی ها نمایانگر ترجیحات یک مصرف کننده و دریافت افزایش توجه از کسب و کارها می باشد. مطابق با آن برخی سیستم های درجه بندی برای کاربران فرصتی را برای درجه بندی گزینه های براساس معیارهای چندگانه فراهم می سازد که می تواند تا حد زیادی اطلاعات درجه بندی را غنی سازد.

همه داده هایی که قبلا اشاره گردید می تواند نقش مهمی را در سیستم پیشنهاد دهنده ایفا کند اگر به طور موثری استفاده گردد. ولیکن، طبق توضیح بخش یک، فیلترسازی مبتنی بر همکاری نیازی به هیچ اطلاعاتی درباره کاربران (داده های مشخصات فردی) و گزینه ها (داده های تولید) ندارد، بر فیدبک کاربر از جمله فیدبک اشکار (درجه بندی کاربر) و فیدبک تلویحی (رفتار کاربر) متمرکز می باشند. ویژگی های کلیدی این دو نوع داده ها در جدول ۱ خلاصه سازی شده است.

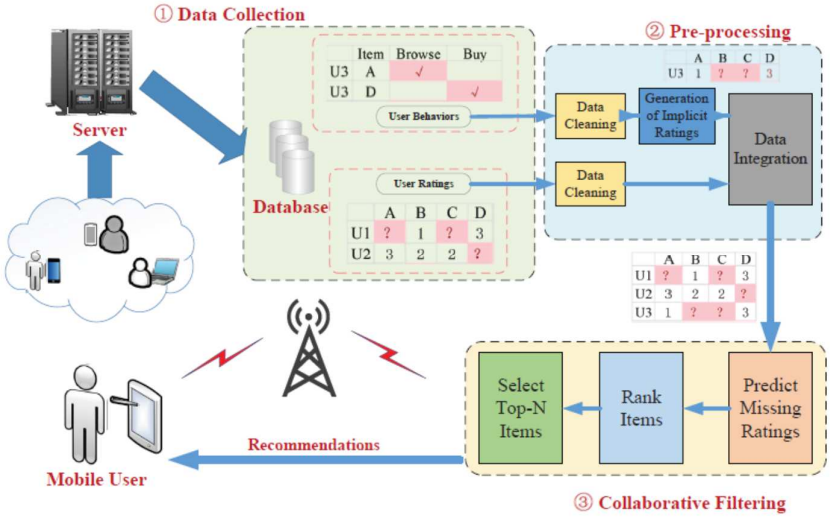
B-پیش پردازش

با پیشرفت اینترنت موبایل، داده های جمع اوری شده معمولا در فرمت های مختلف به دلیل تنوع تجهیزات کاربر و ناهمگنی شبکه ها می باشند. از اینرو، پیش پردازش داده ها یک بخش ضروری سیستم های توصیه کننده می باشد، که مسئول تضمین داده های ورودی فیلترسازی مبتنی بر همکاری برای تکمیل و پایایی می باشد. پیش پردازش معمولا به سه مرحله ذیل تقسیم بندی می شود.

۱) تمیزسازی داده ها: داده های خام نمی تواند مستقیما به دلیل حضور داده های کثیف که می تواند با قصورات تجهیزات احتمالی یا خطاهای انتقال رخ داده باشد، استفاده بشود. نسبت خطا خیلی بالا می شود بویژه زمانی که کاربران در سرعت بالا حرکت کنند. بعلاوه، برخی مصرف کنندگان ممکن است گزینه ها را به طور اختیاری درجه بندی کنند مانند اینکه به همه گزینه های بالاترین درجه بندی را برای صرفه جویی در زمان بدهند که احتمالا پایایی اطلاعات درجه بندی را به طور کلی پایین می اورد. الگوریتم های شناسایی جداکننده ویژه می تواند این مسائل را تا حد زیادی تسکین دهد. برای مثال، بعد از انتخاب بخشی از درجه بندی ها تحت عنوان داده های اموزشی و تعیین یک مدل طبقه بندی کننده براساس الگوریتم های یادگیری ماشینی، جداکننده ها می توانند با صحت رضایت بخشی حذف شوند.

۲) نسل درجه بندی تلویحی: اغلب سیستم های پیشنهاددهنده CF صرفا با درجه بندی کاربر آشکار به شکل اطلاعات ارزشمند رفتار می کنند. ولیکن، یک نسبت بزرگی از کاربران همیشه گزینه هایی را درجه بندی نمی کنند که قبلا مصرف کرده بودند که منجر به مسئله پراکندگی داده ها می شود. به یمن مشتریان موبایل که به طور گسترده بکار بسته شده است ، رفتارهای کاربر مخصوص جمع اوری شده و به شکل توده ای با ارزش احتمالی زیادی ذخیره سازی شده که ممکن است کلیدی برای کم رنگ کردن این مشکل بشود. برای نمونه، سیستم های پیشنهاددهنده حجم های زیادی از درجه بندی های کاربر و رفتارهای کاربر را به شکل مجموعه اموزشی جمع اوری کرده و بعد الگوریتم های یادگیری ماشینی خاصی را روی آن اجرا می کند مانند شبکه عصبی یا درخت تصمیم به نحوی که یک مدل پیشگویی را می سازد که می تواند رفتارهای کاربر را به شکل درجه بندی های تلویحی تغییرشکل بدهد طبق شکل ۲. حجم داده های درجه بندی می تواند تا حد زیادی به این طریق افزایش یابد.

۳) یکپارچه سازی داده ها: هر دو داده های درجه بندی اشکار و تلویحی به شکل یک ماتریس به نام ماتریس درجه بندی طبق شکل ۳ ترکیب شده است. مشهودا، همچنان یک تعداد زیادی عناصر از دست رفته در این ماتریس وجود دارد که باید از طریق فیلترسازی مبتنی برهمکاری پر شود.



شکل ۱-چارچوب سیستم پیشنهاددهنده CF

جدول ۱-ویژگی های کلیدی درجه بندی کاربر و رفتار کاربر

ویژگی	رفتار کاربر	درجه بندی کاربر
اندازه داده ها	بزرگ	کوچک
غیرساختاری	اساسا نیمه ساختاری یا غیرساختاری، معمولا در فایلهای گزارشات ذخیره سازی می شود.	داده های ساختاری، که می تواند به سهولت با یک ماتریس نمایش داده شود.
پوشش	وسیع، تقریبا همه کاربران ثبت می شوند.	محدود، تنها یک بخش از کاربران عادت درجه بندی یک گزینه را بعد از استفاده از آن دارد.
عینی/ذهنی	عینی	ذهنی
آسان/سخت برای استفاده	مشکل، الگوریتم های خاصی نیاز است تا مقدار احتمالی را بررسی کند.	آسان، داده ها می تواند مستقیما ورودی سیستم های پیشنهاد دهنده CF باشند.
قابلیت اتکا	به دلیل الگوریتم های داده کاوی و مقدار داده های آموزشی بی ثبات است.	بالا، که بازتاب ترجیح کاربر طبق یک گزینه خاص است.

شکل ۲-ایجاد درجه بندی های تلویحی براساس رفتارهای کاربر

C-اصول سنجش فیلترسازی مبتنی بر همکاری

روشهای عمومی CF شامل پیشگویی مقادیر از دست رفته، رتبه بندی گزینه ها و انتخاب گزینه های N تعداد رده اول می باشد. چون ماتریس درجه بندی ناقص می باشد، وظیفه اصلی فیلترسازی مبتنی بر همکاری پیشگویی این عناصر مفقود براساس داده های معین می باشد. بعد از آن، گزینه ها طبق درجه بندی های پیشگویی شده رتبه بندی شده و N تعداد اول آنها به شکل پیشنهاداتی انتخاب می شوند. وقتی یک سیستم پیشنهاد دهنده برقرار می شود، چالش دیگر این است که چگونه عملکردش را ارزیابی کنیم. اصول سنجش سیستم های پیشنهاددهنده به سه دسته بندی تقسیم می شود، یعنی:

۱) اصول سنجش صحت پیشگویان: برای تخمین صحت پیشگویی، پایگاه داده کامل به یک مجموعه اموزشی و یک مجموعه تست تقسیم بندی می شود.

	A	B	C	D
U1	?	1	?	3
U2	3	2	2	?
U3	2	?	1	3
U4	4	4	2	?

شکل ۳-یک ماتریس درجه بندی با مقادیر مفقود

مجموعه اموزشی برای ایجاد پیشگویی هایی استفاده می شود درحالیکه مجموعه تست مسئول ارزیابی صحت پیشگویی می باشد. دو اصول سنجش صحت پیشگویانه به طور گسترده ای در فیلترسازی مبتنی بر همکاری بکار بسته می شود یعنی خطای مطلق میانگین MAE و خطای ریشه میانگین مربع RMSE که در معادله ۱ و ۲ معین می شود.

$$MAE = \frac{\sum_{(u,i)\in R_{test}} |R_{u,i} - R'_{u,i}|}{|R_{test}|}, \qquad (1)$$

$$RMSE = \sqrt{\frac{\sum_{(u,i)\in R_{test}} |R_{u,i} - R'_{u,i}|}{|R_{test}|}}, \qquad (2)$$

که در آن $|R_{test}|$ نمایانگر تعداد درجه بندی در مجموعه تست می باشد. $R_{u,i}$ درجه بندی پیشگویی شده برای کاربر u روی گزینه i و $R'_{u,i}$ درجه بندی حقیقی در مجموعه تست می باشد. یک MAE یا RMSE پایین تر نمایانگر یک صحت پیشگویانه بالاتری می باشد.

۲) اصول سنجش صحت طبقه بندی: هرچند اصول سنجش فوق کارامد است و درک راحتی دارد، کاربران ممکن است درباره رقم دقیق درجه بندی پیشگویی شده دقتی نداشته باشند. آنچه تنها به آنها مربوط می شود این است که آیا پیشنهادات به علایق آنها مربوط است یا خیر. تحت این شرایط، یک راه حل عملی قرار است درجه بندی را به مقیاس باینری با انتخاب حداستانه به طور مناسب تغییر شکل بدهد. برای مثال، اگر درجه بندی از ۱الی ۵ بشود، گزینه های درجه بندی شده بیش از ۴ می تواند به شکل مرتبط نگاشته شود و سایرین نامربوط هستند. به این شیوه، هر دو فهرست پیشنهادات و مجموعه تست به دو بخش تقسیم بندی می شود. سه اصول سنجش صحت طبقه بندی به طور وسیعی برای ارزیابی مرتبط سازی بین پیشنهادات و علایق کاربر استفاده می شود یعنی فراخوانی، دقت و نمره F-۱.

فراخوانی به شکل نسبت بین تعداد گزینه های مرتبط در فهرست پیشنهاد و در مجموعه تست تعریف می کنند.

$$recall = \frac{\sum_u L(N,u)}{\sum_u L(u)}, \qquad (3)$$

که در آن $L(N,u)$ و $L(u)$ نمایانگر گزینه های مرتبط برای کاربر u در فهرست پیشنهادات و مجموعه تست به ترتیب می شود. حد بالایی فراخوانی برابر ۱ است که به معنای آنست که همه گزینه های مرتبط در مجموعه تست می تواند در فهرست پیشنهادات یافت شود.

دقت را به شکل درصد گزینه های مرتبط در فهرست پیشنهادات تعریف می کنند.

$$precision = \frac{\sum_u L(N,u)}{UN}, \qquad (4)$$

که در آن U تعداد کاربران می باشد. حد بالایی دقت برابر ۱ است، که به این معناست که همه گزینه های فهرست پیشنهادات مرتبط می باشد.

گاهی اوقات این دو اصول سنجش می تواند با یکدیگر در تناقض باشد. برای مثال، اگر سیستم همه گزینه های پایگاه داده ها را به کاربران به شکل پیشنهاداتی ارائه بدهد، دقت می تواند خیلی پایین باشد حین اینکه فراخوانی تا ۱ هم می رسد. ولیکن، اگر سیستم تنها یک گزینه را به کاربر پیشنهاد نماید و این گزینه دقیقا مرتبط باشد، آنگاه دقت ۱ می شود حین اینکه فراخوانی به طور غیرمحتمل بالا می باشد. برای سنتز دو اصول سنجش، نمره F۱ مطرح می شود که به شکل میانگین هارمونیک بین فراخوانی و دقت تعریف می شود.

$$F_1 = \frac{2 * recall * precision}{recall + precision}. \qquad (5)$$

۳)سنجش تنوع: اخیرا، محققان به شدت آگاه شده اند که رهگیری صرف افزایش صحت پیشگویانه می تواند تنوع پیشنهادات را کاهش بدهد. با اینحساب، سنجش تنوع لازم است که حین ایجاد پیشنهادات درنظرگرفته شده و می تواند به دو قسمت تقسیم بندی شود یعنی شباهت درون فهرست و شباهت بین فهرست. شباهت درون فهرست S_u^i به اندازه گیری شباهت بین هر گزینه در فهرست پیشنهادات برای کاربر u می پردازد که به ترتیب ذیل محاسبه می شود:

$$S_{\text{intra}}^u = \frac{2}{n(n-1)} \sum_{i,j \in L_u, i \neq j} Sim(i,j), \qquad (6)$$

که در آن L_u فهرست پیشنهادات برای کاربر u می باشد. n تعداد پیشنهادات است. $Sim(i,j)$ نمایانگر شباهت بین گزینه i و j می باشد. جزئیات محاسبه شباهت در بخش بعدی ارائه شده است.

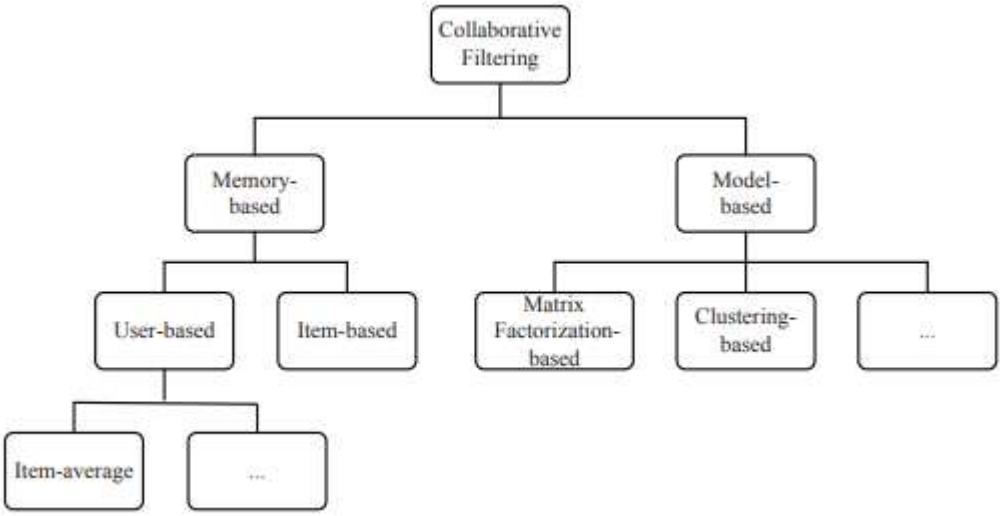
شباهت بین فهرستی برای سنجش شباهت میان پیشنهادات کاربر u و سایر کاربران استفاده شده است که به ترتیب ذیل تعریف می شود:

$$S_{\text{inter}} = \frac{2}{n(n-1)} \sum_{u,v \in U, u \neq v} \frac{|L_u \cap L_v|}{|L_u|}, \qquad (7)$$

که در آن n تعداد پیشنهادات برای کاربر u می باشد. L_u و L_v فهرست پیشنهادات برای کاربر u و v می باشد. مشخص است که یک شباهت بالاتر داخل فهرست یا بین فهرست نشان دهنده یک گرایش پایین تر فهرست پیشنهادات می باشد.

۳-الگوریتم های معمولی CF

طبق تصویر ۴، الگوریتم های CF می توانند به طور تقریبی به دو دسته بندی تقسیم گردند یعنی CF مبتنی بر حافظه و CF مبتنی بر مدل. الگوریتم های CF مبتنی بر حافظه مستقیما از حجم های داده های سابقه ای برای پیشگویی درجه بندی روی گزینه هدف و فراهم سازی پیشنهادات برای کاربر فعال استفاده می کنند. هرگاه یک کار پیشنهاد اجرا شود، این سیستم نیازمند بارگیری همه داده ها به حافظه و اجرای الگوریتم های خاصی روی آنها می باشد. به طور متفاوتی، CF مبتنی بر مدل می تواند از برخی روشهای داده کاوی برای برقراری یک مدل پیشگویی مبتنی بر داده های معلوم استفاده نماید. وقتی مدل کسب گردید، نیازی به داده های خام دیگری حین پیشنهاددهی نیست.



شکل ۴-طبقه بندی الگوریتم های فیلترسازی مبتنی بر همکاری

A) CF مبتنی بر حافظه

CF مبتنی بر حافظه به طور متداولی در سیستم های پیشنهاد دهنده به یمن اجرای با کاریی بالا و اسانش استفاده می شوند. عملکرد CF مبتنی بر حافظه معمولا طبق صحت و تنوع رضایت بخش بوده است. براساس همه درجه بندی های پایگاه داده ها، سیستم پیشنهاد دهنده مجاوران را برای کاربر یا گزینه خاصی می یابد و مقدار پیشگویی شده را برای درجه بندی های نامعلوم محاسبه می کند.

۱)طبقه بندی: الگوریتم های CF مبتنی بر حافظه می تواند به دو نوع تقسیم بندی شود: CF مبتنی بر کاربر و CF مبتنی بر گزینه. CF مبتنی بر کاربر رابطه میان ردیف ها را در ماتریس درجه بندی بررسی می کند در حالیکه CF مبتنی بر گزینه بر رابطه میان ستونها متمرکز است.

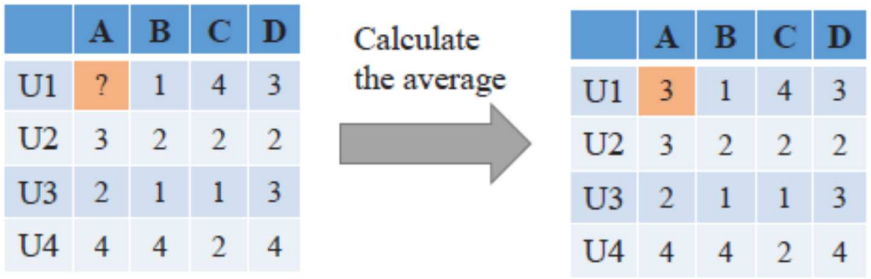
a) CF مبتنی بر کاربر: CF مبتنی بر کاربر ابتدا شباهت میان کاربر فعال و سایرکاربران را محاسبه می کند. سنجشهای شباهت مشترک در CF شامل کسینوس خالص، کسینوس تنظیم شده و ضریب همبستگی پیرسون می باشد. کاربران با شباهت های بالا به شکل مجاوران کاربر فعال انتخاب می شوند. بعد سیستم از درجه بندی های مجاوران روی یک گزینه مخصوص برای محاسبه متوسط توزین شده استفاده می کند که به شکل درجه بندی پیشگویی شده تلقی می شود، و شباهت های مربوطه را به شکل توزین هایی تلقی می کند. دست کم،

پیشنهاددهنده همه گزینه ها را طبق درجه بندی های پیشگویی شده شان رتبه بندی می کند و N گزینه اول را به شکل پیشنهاداتی انتخاب می کند. یک چالش که CF مبتنی بر کاربر باید با آن روبرو شود مسئله قابلیت مقیاس بندی سات. برای برخی وب سایت های ویدئویی مشهور با میلیونها کاربر ثبت نامی، محاسبه شباهت های میان همه کاربران و انتخاب مجاوران در زمان واقعی اجرای مشکلی دارد. در نتیجه، CF مبتنی بر کاربر مناسب تر است زمانی که کاربران به تعداد زیاد نبوده و گروه کاربر نسبتا با ثبات باشد.

الگوریتم گزینه-متوسط یک مورد خاص CF مبتنی بر کاربر است که همه کاربران را به شکل مجاوران با توزین های معادل طبق شکل ۵ انتخاب می کنند.

b) CF مبتنی بر گزینه:

برخلاف CF مبتنی بر کاربر، CF مبتنی بر گزینه متمرکز بر شباهت ها در میان گزینه ها می باشد. این امر براساس این فرضیه است که گزینه های دارای درجه بندی های کاربر مشابه احتمالا از انواع مشابهی هستند. با اینحساب، شباهت ها در میان گزینه ها ابتدا با استفاده از سنجشهای شباهت یکسانی با CF مبتنی بر کاربر محاسبه می شود



شکل ۵-پیشگویی درجه بندی با استفاده از الگوریتم گزینه-متوسط

بعد از انتخاب مجاوران برای گزینه هدف و محاسبه متوسط توزین شده، درجه بندی پیشگویی شده روی این گزینه بدست می اید. اسان است که درک کنیم زمانی که گزینه ها خیلی زیاد می شوندو مرتب هم تغییر می کنند، مسئله قابلیت مقیاس پذیری نیز اجتناب دشواری دارد. یک مقایسه میان CF مبتنی بر کاربر و CF مبتنی بر گزینه در جدول ۲ نشان داده شده است.

۲)عملیات کلی: شکل ۶ پنج عملیات کلی CF مبتنی بر حافظه را ارائه می دهد که به تحلیل مفصل به ترتیب ذیل نیاز دارد:

-**مرحله ۱**-محاسبه شباهت

محاسبه شباهت میان کاربران و گزینه ها یک مرحله حیاتی در CF محسوب می شود. مقادیر زیادی از اندازه گیری های شباهت تا حد زیادی در CF مبتنی بر حافظه استفاده شده است، برای مثال:

a) اندازه گیری های شباهت پایه:

-شباهت کسینوسی خالص:

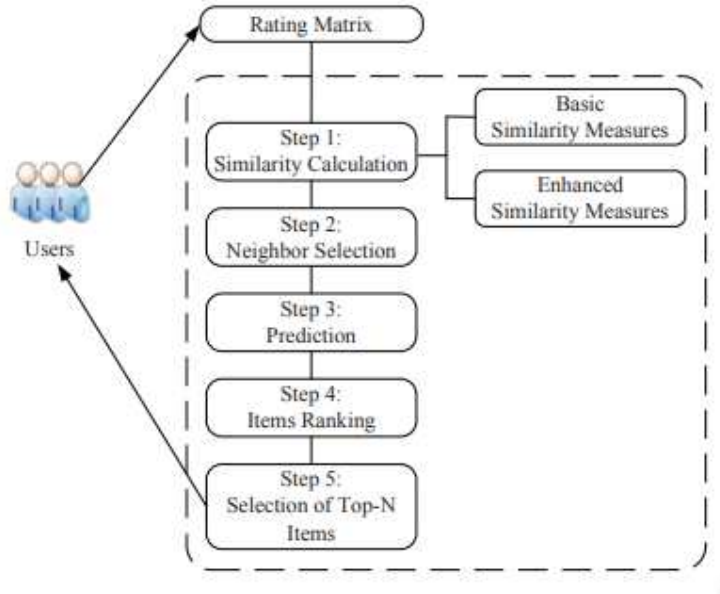
شباهت کسینوس خالص مقدار کسینوس زاویه میان دو بردار را اندازه گیری می کند. برای CF مبتنی بر کاربر، کاربران با ردیفهای ماتریس درجه بندی با مجموعه مقادیر مفقوده تا ۰ نمایش داده می شوند. شباهت کسینوس خالص بین بردارهای کاربر به ترتیب ذیل محاسبه می شود:

$$\omega_{ij}^{u} = cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{||\vec{i}||_2 * ||\vec{j}||_2}. \qquad (8)$$

که در آن i و j نشاندهنده دو بردار کاربر می باشد. برای CF مبتنی بر گزینه، گزینه ها با ستونهای نمایش داده می شوند.

جدول ۲-مقایسه میان CF مبتنی بر کاربر و CF مبتنی بر گزینه

شرح برنامه های اپلیکیشن	شباهت	دسته جات CF
-تعداد گزینه ها بزرگتر از مال کاربر است. -کاربران مرتب تغییر نمی کنند.	شباهت کاربر با کاربر	CF مبتنی بر کاربر
-تعداد کاربران بزرگتر از مال گزینه هاست. -گزینه ها مرتب تغییر نمی کند.	شباهت گزینه با گزینه	CF مبتنی بر گزینه



شکل ۶-روشهای اصلی CF مبتنی بر حافظه

ماتریس درجه بندی و ماتریس بین آنها به ترتیب ذیل تعریف می شود:

$$\omega_{kl}^{i} = cos(\vec{k}, \vec{l}) = \frac{\vec{k} \cdot \vec{l}}{||\vec{k}||_2 * ||\vec{l}||_2}. \qquad (9)$$

که در آن k و l نمایانگر دو بردار گزینه می باشند.

-شباهت کسینوسی تنظیم شده:

چون مقیاسهای درجه بندی در میان کاربران متفاوت می باشد، همان درجه بندی دو کاربر به معنای درجه بندی مورد علاقه نمی باشد. این مسئله به دلیل شباهت کسینوسی خالص نادیده گرفته شده است. بعلاوه، تنظیم درجه بندی های مفقوده به ۰ بنا به فرض بیش و کم غیرمنطقی می باشد. شباهت کسینوسی تنظیم شده باعث تصحیح این نواقص با کسر متوسط درجه بندی کاربر و استفاده از گزینه های با درجه بندی همزمان برای تعیین بردار می شود. گزینه های درجه بندی شده همزمان گزینه هایی هستند که با هر دو کاربر i و کاربر j درجه بندی می شوند. شباهت کسینوسی تنظیم شده برای CF مبتنی بر کاربر به ترتیب ذیل معین می شود:

$$\omega_{ij}^{u} = \frac{\sum_{k \in K}(r_{i,k} - \bar{r_i})(r_{j,k} - \bar{r_j})}{\sqrt{\sum_{k \in K}(r_{i,k} - \bar{r_i})^2}\sqrt{\sum_{k \in K}(r_{j,k} - \bar{r_j})^2}}, \qquad (10)$$

که در آن K مجموعه گزینه های با درجه بندی همزمان می باشد. $r_{i,k}$ و $r_{j,k}$ دو درجه بندی درباره گزینه k از کاربر i و j می باشد. $\bar{r}_i$ و $\bar{r}_j$ درجه بندی های متوسط کاربر i و کاربر j می باشد. برای الگوریتم مبتنی بر گزینه، کاربران با درجه بندی همزمان انتخاب شده اند تا شباهت بین گزینه ها را محاسبه کنند که در معادله ۱۱ نشان داده شده است.

$$\omega_{kl}^{i} = \frac{\sum_{u \in U}(r_{u,k}-\bar{r_u})(r_{u,l}-\bar{r_u})}{\sqrt{\sum_{u \in U}(r_{u,k}-\bar{r_u})^2}\sqrt{\sum_{u \in U}(r_{u,l}-\bar{r_u})^2}}, \quad (11)$$

که در آن U مجموعه کاربرانی است که هر دو گزینه k و l را رتبه بندی کرده اند. $r_{u,k}$ و $r_{u,l}$ دو درجه بندی از کاربر u درباره گزینه k و l می باشد. $\bar{r}_u$ متوسط درجه بندی کاربر u می باشد.

-ضریب همبستگی همبستگی پیرسون:

ضریب همبستگی پیرسون PCC بازتاب درجه همبستگی خطی بین دو متغیر می باشد. همانند کسینوس تنظیم شده، PCC گزینه ها یا کاربران با درجه بندی همزمان را برای محاسبه شباهت ها انتخاب می کند. ضرایب همبستگی پیرسون بین دو کاربر یا دو گزینه به ترتیب با معادله ۱۲ و ۱۳ معین می شود.

$$\omega_{ij}^{u} = \frac{\sum_{k \in K}(r_{i,k}-\bar{r_i})(r_{j,k}-\bar{r_j})}{\sqrt{\sum_{k \in K}(r_{i,k}-\bar{r_i})^2}\sqrt{\sum_{k \in K}(r_{j,k}-\bar{r_j})^2}}, \quad (12)$$

$$\omega_{kl}^{i} = \frac{\sum_{u \in U}(r_{u,k}-\bar{r_k})(r_{u,l}-\bar{r_l})}{\sqrt{\sum_{u \in U}(r_{u,k}-\bar{r_k})^2}\sqrt{\sum_{u \in U}(r_{u,l}-\bar{r_l})^2}}. \quad (13)$$

می توان مشاهده کرد که برای CF مبتنی بر کاربر، PCC همانند کسینوس تنظیم شده است، در حالیکه برای CF مبتنی بر گزینه، آنها اندکی متفاوت می باشند. کاستگرهای درجه بندی ها همان متوسط کاربران و گزینه ها به ترتیب می باشند.

b) سنجشهای شباهت تقویت شده:

-شباهت مبتنی بر مجموعه:

شباهت احتمالا تخمین بیش از حد با استفاده از کسینوس تنظیم شده یا PCC می باشد هنگامی که تعداد گزینه های رتبه بندی شده همزمان خیلی کوچک باشد. یک شباهت مبتنی بر مجموعه برای رفع این مشکل مطرح شده است. برای CF مبتنی بر کاربر، شباهت تقویت شده بین کاربر i و کاربر j به ترتیب ذیل تعریف شده است:

$$\omega'^{u}_{ij} = \frac{2|K_i \cap K_j|}{|K_i| + |K_j|}\omega^{u}_{ij}, \qquad (14)$$

که در آن ω^{u}_{ij} شباهت پایه نظیر PCC می باشد. $|K_i|$ و $|K_j|$ تعداد گزینه های درجه بندی شده توسط کاربر i و کاربر j به ترتیب می باشند. $|K_i \cap K_j|$ نشان دهنده تعداد گزینه های با درجه بندی همزمان کاربر i و کاربر j می باشد. اگر گزینه های درجه بندی شده همزمان خیلی معدود باشند، شباهت پایه تا حد زیادی طبق عامل تباهی کاهش خواهد یافت. بعلاوه، شباهت تقویت شده بین گزینه k و گزینه l به شکل ذیل تعریف می شود:

$$\omega'^{i}_{kl} = \frac{2|U_k \cap U_l|}{|U_k| + |U_l|}\omega^{i}_{kl}. \qquad (15)$$

-شباهت آگاه از زمان:

حین اینکه علایق کاربران می تواند طی زمان تغییر یابد، یک تعداد رو به رشدی از سیستم های پیشنهاد دهنده تاثیر زمان را روی صحت پیشگویانه درک کرده اند. رفرانس ۵۰ یک الگوریتم CF اگاه از زمان را مطرح کرده که اثر زمان را روی محاسبه شباهت درنظر گرفته است. از یک سو، درجه بندی ها با مهرهای زمانی مشابه بیشتر در شباهت کاربر نقش دارند و این نقش به شکل ذیل نشان داده می شود:

$$f_1(t_{ik}, t_{jk}) = e^{-\alpha|t_{ik}-t_{jk}|}, \qquad (16)$$

که در آن t_{ik} و t_{jk} مهرهای زمانی هستند زمانی که کاربر i و کاربر j درجه بندی گزینه k را انجام می دهند. آلفا یک ثابت غیرمنفی است، که تصمیم می گیرد که به چه سرعتی f_1 با افزایش $|t_{ik} - t_{jk}|$ کاهش می یابد. از سوی دیگر، درجه بندی های تازه تر بیشتر در شباهت کاربر نقش دارند. به عبارت دیگر، اگر دو

کاربر همان گزینه را مدتها قبل درجه بندی کرده باشند، درجه بندی آنها اهمیت کمتری روی پیشگویی در زمان کنونی می یابد. این نقش به ترتیب ذیل تعریف می شود:

$$f_2(t_{ik}, t_{jk}) = e^{-\beta|t_{current}-(t_{ik}+t_{jk})/2}, \qquad (17)$$

که در آن $t_{current}$ زمانی است که پیشنهاد اجرا می شود. براساس تحلیل فوق، شباهت آگاهی از زمان بین کاربر i و j با معادله ۱۸ تعریف می شود:

$$\omega_{ij} = \frac{\sum_{k\in K}(r_{i,k}-\bar{r_i})(r_{j,k}-\bar{r_j})f_1(t_{ik},t_{jk})f_2(t_{ik},t_{jk})}{\sqrt{\sum_{k\in K}(r_{i,k}-\bar{r_i})^2}\sqrt{\sum_{k\in K}(r_{j,k}-\bar{r_j})^2}}. \quad (18)$$

-**مرحله دوم**: انتخاب مجاور

صحت پیشگویی وقتی برخی کاربران نامشابه در مجاورت مشغول باشند، کاهش خواهد یافت. از اینرو مجاوران کاربر فعال باید به دقت با روشهای خاصی انتخاب شوند. الگوریتم N رده اول معمولی به انتخاب N تا از مشابه ترین مجاوران برایا نجام پیشگویی ها می پردازد که این پدیده را نادیده می گیرد که برخی کاربران ممکن است تعداد محدودی مجاوران کمتر از N را داشته باشند. رفرانس ۳۰ مطرح کرده است که یک مجاور کاندید که شباهت آن کمتر از صفر است باید از مجموعه N رده اول حذف شود که به ترتیب ذیل تعریف می شود:

$$S(i) = \{i_a | i_a \in N(i), \omega^u_{ii_a} > 0, i_a \neq i\}, \qquad (19)$$

که در آن $N(i)$ مجموعه ای از کاربران مشابه N رده اول می باشد. به طور متناظر، مجاوران گزینه k به ترتیب ذیل تعریف می شود:

$$S(k) = \{k_a | k_a \in N(k), \omega^i_{kk_a} > 0, k_a \neq k\}, \qquad (20)$$

که در آن $N(k)$ تعداد N گزینه مشابه اول فهرست می باشد.

بعلاوه، افراد در کشورها یا نواحی مختلف به احتمال بیشتری دارای ترجیحات متفاوت می باشند. از اینرو موقعیت های کاربر باید درنظر گرفته شود حین اینکه انتخاب مجاوران برای کاربر فعال صورت می گیرد. به یمن پیشرفت اینترنت موبایل، اطلاعات موقعیت جغرافیایی می تواند از طریق یا مشتری موبایل یا ادرس IP بدست آید، و به

سرور برای تحلیل بیشتر انتقال یابد. معمولا، کاربران می توانند ابتدا به قسمت بندی های متعدد بسته به موقعیت آنها تقسیم بندی شود و کاربران نزدیک از لحاظ فضایی اولویتی در انتخاب مجاوران دارند. برای نمونه، رفرانس ۳۲ یک مدل سه ردیفی را از روابط فضایی مطرح کرده است یعنی سیستم خودمختار AS یکسان، کشور یکسان و سایر موارد. کاربرانی که به حداستانه شباهت می رسند و رابطه فضایی نزدیکتری دارند قبلا به عنوان مجاوران کاربر فعال انتخاب شده اند.

مرحله ۳-پیشگویی

با شباهت ها و درجه بندی های مجاوران، سیستم پیشنهاددهنده متوسط توزین شده را به شکل پیشگویی محاسبه می کند. برای CF مبتنی بر کاربر، درجه بندی پیشگویی شده کاربر i روی گزینه k به ترتیب ذیل محاسبه می شود:

$$P(i,k) = \bar{r}_i + \frac{\sum_{i_a \in S(i)} \omega^u_{ii_a} (r_{i_a,k} - \bar{r}_{i_a})}{\sum_{i_a \in S(i)} \omega^u_{ii_a}}, \qquad (21)$$

که در آن $S(i)$ مجموعه مجاوران کاربر i می باشد. $\omega^u_{ii_a}$ شباهت میان کاربر i و i_a می باشد. برای CF مبتنی بر گزینه، درجه بندی پیشگویی شده کاربر i روی گزینه k در معادله ۲۲ نشان داده شده است.

$$P(i,k) = \bar{r}_k + \frac{\sum_{k_a \in S(k)} \omega^i_{kk_a} (r_{i,k_a} - \bar{r}_{k_a})}{\sum_{k_a \in S(k)} \omega^i_{kk_a}}, \qquad (22)$$

که در آن $S(k)$ مجموعه مجاوران برای گزینه k می باشد. $\omega^u_{kk_a}$ شباهت میان گزینه k و k_a می باشد.

-**مرحله ۴**: رتبه بندی گزینه ها

وقتی پیشگویی ها بدست آمد، سیستم پیشنهاد دهنده باید همه گزینه ها را طبق رتبه بندی های پیشگویی شده شان رتبه بندی نماید. برای بهبود تنوع پیشنهادات، برخی سیستم های پیشنهاددهنده نیز مورد توجه قرار گرفته اند. یک گزینه دارای مقدار پیشگویی شده بزرگتر و شهرت کمتر بنا به فرض رتبه بندی بالاتری دارند.

-**مرحله ۵**: انتخاب N گزینه اول فهرست.

بعد از رتبه بندی همه گزینه های کاندیدا، N گزینه اول فهرست انها برای کاربری فراهم شده که در آن N یک پارامتر مورد نیاز برای تنظیم قبل از کار پیشنهادات می باشد.

CF-B مبتنی بر مدل

هر چند CF مبتنی بر حافظه در پیشگویی موثر درجه بندی های مفقوده و ارائه توصیه ها مفید می باشد، همچنان یک محدودیت های معدودی وجود دارد. برای نمونه، هر گاه یک کار توصیه اجرا گردد، سیستم مجبور است که همه درجه بندی ها را به حافظه بارگذاری کند و الگوریتم خاصی را براساس پایگاه داده کامل اجرا نماید. CF مبتنی بر حافظه که با منابع ذخیره سازی و محاسبه محدود شده است، می تواند اغلب کاملا زمانبر باشد. از اینرو، سیستم پیشنهاد دهنده که می تواند گزینه های مناسبی را با مصرف زمان قابل قبول فراهم سازد به شدت دلخواه و مطلوب می باشد. الگوریتم های CF مبتنی بر مدل برای رفع این مسائل طراحی شده اند که اصل کلی اش استفاده از یادگیری ماشینی یا روشهای داده کاوی برای تعیین مدلهای پیشگویی به شکل افلاین می باشد. براساس این مدلها، درجه بندی های مفقوده می تواند به طور موثری پیشگویی شود. الگوریتم های مبتنی بر مدل معمولی شامل الگوریتم های مبتنی بر فاکتورگیری ماتریس، الگوریتم های مبتنی بر خوشه گیری و غیره می باشد.

۱) الگوریتم های مبتنی بر فاکتورگیری: پراکندگی ماتریس درجه بندی همیشه چالش اصلی است که عملکرد فیلتربندی مبتنی بر همکاری را محدود می سازد. علت این مسئله این است که ابعاد بردار کاربران یا گزینه ها همیشه خیلی بزرگ است. الگوریتم عامل گیری ماتریس MF که یکی از روشهای یادگیری بدون نظارت می باشد، می تواند نقشی را در کاهش بعدگرایی و سرانجام رفع پراکندگی داده ها ایفا نماید. روشهای اصلی الگوریتم های CF مبتنی بر فاکتورگیری ماتریس در شکل ۸ نشان داده شده است.

a) عملیات ۱-مدلسازی ویژگی نهفته: ماتریس درجه بندی معمولا حاوی برخی ویژگی های نهفته می باشد که می تواند برای توضیح مشخصات کاربران و گزینه ها به طور اخص تر استفاده گردد. با گرفتن ویدئوهایی برای مثال، ویژگی های نهفته می تواند سبک های ویدئوها مانند فیلم کمدی، تراژدی، و غیره باشد. طبق تصویر ۷، بردار ویژگی کاربر p_u نشان می دهد که کاربر u چقدر به هر ویژگی علاقمند است و بردار ویژگی گزینه q_i درجه هر ویژگی را برای گزینه i اندازه گیری می کند. براساس این دو بردار، درجه بندی کاربر u روی گزینه i می تواند با معادله ۲۳ محاسبه گردد.

$$r_{u,i} = p_u^T q_i. \tag{23}$$

b) عملیات ۲-تعیین هدف بهینه سازی: آسان است که درک کنیم وقتی ماتریس ویژگی کاربران و گزینه ها برقرار می شود، همه درجه بندی های از دست رفته می تواند به راحتی با محاسبه محصول نقطه ای بردارهای ویژگی خاص بدست آید. ولیکن، روشهای عامل گیری ماتریس قدیمی، مانند SVD و PCA در تجزیه ماتریس درجه بندی به دلیل تعداد زیاد مقادیر از دست رفته شکست خورده اند. در واقع، دو ماتریس دلخواه برای رعایت الزاماتی لازم است که محصول نقطه ای میان p_u و q_i به مقدار معین $r_{u,i}$ در ماتریس درجه بندی نزدیک است. در نتیجه می تواند به شکل یک مسئله بهینه سازی با هدف تعریف شده به شکل ذیل مدلسازی گردد:

$$\min_{q^*,p^*} \sum_{(u,i)\in\kappa} \left(r_{u,i} - q_i^T p_u\right)^2 + \lambda \left(\|q_i\|^2 + \|p_u\|^2\right), \tag{24}$$

که در آن κ مجموعه جفت کاربر-گزینه می باشد و $r_{u,i}$ می تواند از مجموعه اموزشی بدست آید. یک پارامتر اضافه λ برای رفع مسئله بیش برازش ارائه شده است.

فرمول بالا گاهی اوقات باید با مسائل سوگیری از سوی کاربران یا گزینه ها روبرو گردد. برای مثال، درجه بندی ها از یک کاربر حیاتی می تواند پایین تر از سایرین باشد. با اینحساب، به متغیرهای اضافی برای درنظرگیری سوگیری نیاز است و تابع عینی جدید به ترتیب ذیل نشان داده شده است:

$$\begin{aligned}\min_{q^*,p^*} \sum_{(u,i)\in\kappa} &\left(r_{u,i} - \mu - b_i - b_u - q_i^T p_u\right)^2 \\ &+\lambda \left(\|q_i\|^2 + \|p_u\|^2 + b_i^2 + b_u^2\right),\end{aligned} \tag{25}$$

که در آن μ متوسط کلیه درجه بندی ها می باشد. b_u تفاوت میان μ و متوسط درجه بندی کاربر u می باشد. b_i تفاوت میان μ و متوسط درجه بندی روی گزینه i می باشد.

درجه بندی ها معمولا طی زمان به دلیل تناوب علایق کاربر یا کاهش شهرت گزینه ها تغییر می کنند. در این خصوص، عوامل زمانی به ملاحظه درآمده و تابع عینی به صورت ذیل بازسازی می شود:

$$\min_{q^*,p^*} \sum_{(u,i)\in\kappa} \left(r_{u,i}(t) - \mu(t) - b_i(t) - b_u(t) - q_i^T p_u(t)\right)^2 + \lambda\left(\|q_i(t)\|^2 + \|p_u(t)\|^2 + b_u(t)^2 + b_i(t)^2\right), \quad (26)$$

$r_{u,i}(t)$ ، $\mu(t)$ ، $b_u(t)$ ، $b_i(t)$ ، $q_i(t)$ و $p_u(t)$ به شکل توابع زمانی تلقی می شوند.

c) عملیات ۳-حل مسئله بهینه سازی:

بسیاری الگوریتم ها مطرح شده تا مسائل بهینه سازی فوق را حل کند و پراستفاده ترین آنها گرادیان تصادفی نزولی و مربعات حداقل متناوب می باشد.

گرادیان تصادفی نزولی SGD یک الگوریتم تکراری است که اصل کلی اش همان روزامدسازی پارامترهای ناشناخته طبق جهت نزولی گرادیان تابع عینی می باشد. برای نمونه، برای حل معادله ۲۴، p_u و q_i به طور تصادفی ابتدا اغاز می شوند. و بعد خطای پیشگویی به ترتیب ذیل محاسبه می شود:

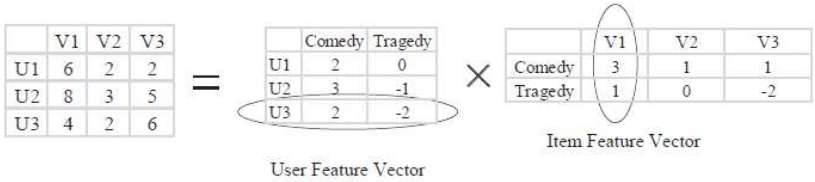
$$e_{u,i} = r_{u,i} - q_i^T p_u. \quad (27)$$

بعد p_u و q_i در جهت مخالف گرادیان اصلاح می شوند:

$$\begin{aligned} q_i &\leftarrow q_i + \alpha(e_{u,i}p_u - \beta q_i), \\ p_u &\leftarrow p_u + \alpha(e_{u,i}q_i - \beta p_u), \end{aligned} \quad (28)$$

که در آن α و β دو ثابتی هستند که می تواند بر میزان همگرایی اثر بگذارند.

جدای SGD، مربعات حداقل متناوب نیز یک روش موثر برای حل و فصل این مسائل می باشد.



شکل ۷-مثالی از فاکتورگیری ماتریس در پیشنهاد ویدیو

برای حل معادله ۲۴، این الگوریتم ابتدا یکی از متغیرها را ثابت کرده و متغیر دیگر را با سروکار داشتن با مسئله مربعات حداقل مورد رسیدگی قرار می دهد. بعد دو متغیر چرخش می یابد، دومی تثبیت شده و اولی هم محاسبه می شود. این عملیات ادامه می یابد تا زمانی که خطای پیشگویی به سمت یک مقدار ثابتی همگرایی یابد. بدین وسیله، SVD می تواند روی یک ماتریس ناقصی اجرا گردد.

۲) الگوریتم های مبتنی بر خوشه گیری: با افزایش مقدار داده ها، محاسبه شباهت ها بین کاربر فعال و همه کاربران دیگر در پایگاه داده ها یک کار خیلی پیچیده ای شده است. به عنوان یکی از پراستفاده ترین روشهای داده کاوی ، خوشه گیری می تواند تا حد زیادی باعث کاهش زمان و منابع محاسباتی طی پیشنهادات بشود. بعد از یک مجموعه عملیات ویژه، داده های ورودی به چندین قسمت تقسیم بندی می شود. اشیای موجود در همان قسمت از شباهت بالاتری با یکدیگر نسبت به اشیای بین قسمت بندیها برخوردارند. براساس خوشه گیری، سیستم های پیشنهاددهنده می تواند گزینه های مناسبی را با قابلیت اتکای بالاتر و پیچیدگی محاسباتی پایین تر فراهم سازد. شکل ۹ روشهای اصلی الگوریتم های مبتنی بر خوشه گیری را در سیستم های پیشنهاددهنده CF رسم کرده است.

a) عملیات ۱-مدلسازی شی خوشه گیری: اولین اولویت خوشه گیری همان پاسخ دهی به سوالات کدام خوشه و روش نمایش آنها می باشد. عموما، هر دو کاربران و گزینه ها می تواند به شکل اشیای خوشه گیری تلقی بشود. وقتی شی تعیین می شود، انواع مدلهای ریاضی را می توان برای ارائه آن بکار برد. بردارهای درجه بندی می تواند مستقیما برای نمایش کاربران و گزینه ها استفاده بشود. مخصوصا، یک گزینه می تواند به شکل بردارهای چندبعدی مدلسازی بشود که ویژگی های آنها درجه بندی های از سوی کاربران می باشد. مشابها، یک کاربر می تواند با همه درجه بندی هایی نمایش داده بشود که برای هر گزینه انجام داده است.

هرچند روش فوق به سهولت درک می شود، همچنان زمانبر بوده چرا که ابعاد بردارها بالا می باشد. برخی اطلاعات اضافی می تواند نقشی حیاتی را در این روش ایفا نماید. برای نمونه، یک گزینه می تواند با تعدادی از کلیدواژه هایی نمایان بشود که عملکردها یا ویژگی هایش را توضیح داده و این اطلاعات می تواند به شکل یک مجموعه مدلسازی شود. گزینه های دارای عناصر مجموعه مشابه احتمالا به شکل مجاوران تلقی می شود. به همان شیوه،

کاربران می توانند با اطلاعات فردی مانند جنسیت، سن و غیره نمایش داده شوند. در نتیجه، بعد مزبور به طور برجسته ای در مقایسه با بردار درجه بندی کاهش یافته است.

اطلاعات توضیحی می تواند به طور موثری نمایانگر کاربران یا گزینه ها باشد. ولیکن این داده ها همیشه قابل دسترسی نیستند. معمولا، تنها داده های درجه بندی شده برای سیستم پیشنهاددهنده در دسترس می باشند. الگوریتم های کاهش بعدگرایی می تواند تحت این شرایط مفید واقع شود. برای مثال، با کمک تحلیل محتوایی اصلی یا PCA روی ماتریس درجه بندی، ویژگی های اصلی را می توان بدست اورد تا تقریبا نمایانگر بردارهای قبلی باشد. بعد از آن، خوشه گیری می تواند روی این بردارها با یک بعد خیلی پایین تری اجرا بشود.

b) عملیات ۲-محاسبه شباهت: وقتی کاربران یا گزینه ها توسط مدلهای ریاضی خاصی نمایش داده می شوند، انتخاب سنجشهای شباهت حیاتی می شود. سنجشهای شباهت پایه مانند فاصله اقلیدسی یا PCC همچنان برای بردارها روایی دارد. بعلاوه، اگر اشیای خوشه گیری یک گروه از مجموعه ها باشد، آمار خاص استفاده شده برای اندازه گیری شباهت میان دو مجموعه به میان می آید. برای مثال، ضریب همبستگی شباهت Jaccard یا JSC، که به شکل اندازه قسمت بینابینی تقسیم بر اندازه اتحادیه تعریف می شود، به طور معمول در سیستم های توصیه کننده مبتنی بر مدل استفاده می شود.

$$\mathrm{Sim}(s_i, s_j) = \frac{|s_i \cap s_j|}{|s_i \cup s_j|}, \qquad (29)$$

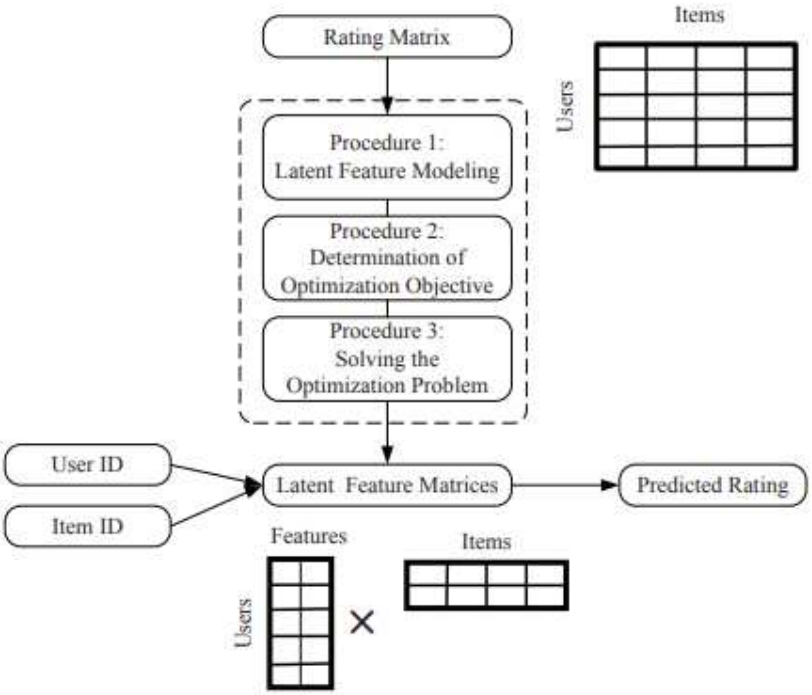
که در آن s_i و s_j دو مجموعه بوده و $|\ |$ نمایانگر اصل مهم مجموعه می باشد.

c) روش ۳-خوشه گیری: بعد از تعیین سنجش شباهت، مرحله بعدی بکارگیری الگوریتم های خوشه گیری خاصی در پایگاه داده ها می باشد. الگوریتم های خوشه گیری می تواند تقریبا به دو بخش تقسیم بندی شود: خوشه گیری نسبی و خوشه گیری سلسله مراتبی. یک نماینده معمول از الگوریتم های خوشه گیری نسبی k میانگین مشهور است که می تواند پایگاه داده کامل را به k قسمت سریعا و با کارایی تقسیم بندی کند. ولیکن، پارامتر پیش تنظیم K اثر مهمی بر نتایجی دارد که تخمین دشواری قبل از خوشه گیری دارد. برای برطرف سازی این مشکل، خوشه گیری سلسله مراتبی مطرح گردیده است که می تواند یک دندروگرام ایجاد کند که سلسله مراتب خوشه گیری را

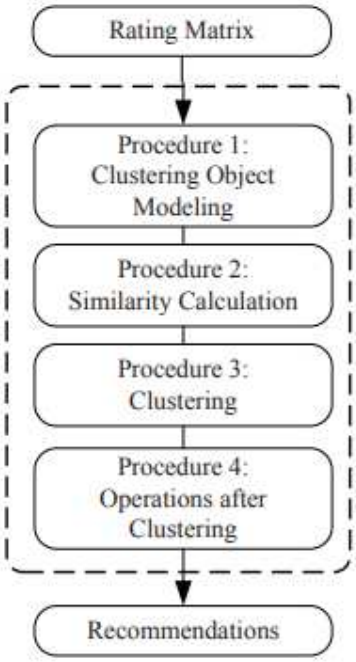
نمایش بدهد. براساس این دندوگرام، انواع نتایج خوشه گیری با تعداد مختلف قسمت بندی ها می تواند به سهولت کسب شود.

d) عملیات ۴-عملیات بعد از خوشه گیری: خوشه گیری اغلب به عنوان مرحله حدواسط CF درنظر گرفته شده و عملیات بیشتر در زمینه نتایج مورد نیاز است. برای مثال، بعد از شناسایی شهرت هر گزینه در خوشه کاربر فعال، برخی گزینه های جذاب می تواند مستقیما به شکل پیشنهاداتی انتخاب شود. بعلاوه، وقتی خوشه کاربر فعال تعیین می شود، کار انتخاب مجاور می تواند صرفا روی کاربران درون خوشه به جای همه کاربران پایگاه داده ها اجرا بشود. در نتیجه، پیچیدگی محاسباتی را می توان به طور برجسته ای کاهش داد.

۳-سایر الگوریتم های مبتنی بر مدل: جدای از الگوریتم های مبتنی بر فاکتورگیری و مبتنی بر خوشه گیری ماتریس، تعداد زیادی مدلهای ریاضی در CF مبتنی بر مدل نیز بکار گرفته شده اند برای مثال شبکه بیزی، پیمایش تصادفی، یادگیری عمیق و غیره. این شیوه ها یک توجه روزافزونی را به دلیل صحت و کارایی بالا مبذول داشته اند. مدل طبقه بندی بیزی ساده را می توان برای پیشگویی درجه بندی کاربر u روی گزینه i براساس توزیع احتمالات درجه بندی های معین بکار بست.



شکل ۸-روش های اصلی الگوریتم های مبتنی بر فاکتورگیری ماتریس



شکل ۹-روش اصلی الگوریتم های مبتنی بر خوشه گیری

به کمک این مدل احتمالات همه درجه بندی های احتمالی را می توان بدست آورد و درجه بندی ها با بالاترین احتمالات به شکل پیشگویی انتخاب شده و به ترتیب ذیل تعریف می شود:

$$\mathrm{R}_p = \underset{r \in RatingSet}{arg\max} \; P(R_r) \prod_n P(X_n = x_n | Y = R_r), \qquad (30)$$

که در آن R_p درجه بندی پیشگویی شده می باشد. X_n نشاندهنده درجه بندی روی گزینه i از سایر کاربران می باشد. R_r نمایانگر درجه بندی کاربر u روی گزینه های دیگر است.

۴-مطالعات موردی یا سیستم های پیشنهاددهنده براساس الگوریتم های CF

در سالهای اخیر، یک تعداد فزاینده ای از افراد ترجیح می دادند که فیلم یا تلویزیون را با استفاده از اپلیکیشن های موبایل تماشا کنند. ازا ینرو، وب سایت های ویدئویی توجه روزافزونی را به خود جلب کردند مانند یوتیوب و نت فلیکس . فیلترسازی مبتنی بر همکاری به این وب سایت ها کمک کرده تا پیشنهادات مطلوبی را برای مصرف کنندگان فراهم کنند مبادا آنها وقت زیادی را صرف ویدئوهای دلخواهشان نمایند. برای توضیح سیستم

پیشنهاددهندهCF به طور اخص تر، دو مطالعه موردی در این بخش براساس یا رفتارهای کاربر یا درجه بندی های کاربر ارائه می شود. آزمایشات انجام شده در این مطالعات موردی با هدف مقایسه عملکردها در میان چندین الگوریتم CF معمولی می باشد و اثر پارامترهای کلیدی را روی MAE نشان می دهد. نتایج ازمایشی نشان داده است که هم درجه بندی های کاربر و هم رفتارهای کاربر می تواند در CFاز طریق عملیات پیش پردازش ویژه ای بکار گرفته شود. بعلاوه، الگوریتم های CF معمولا یک بهبود عظیمی بر صحت پیشگویانه در مقایسه با خط مبنا دارند و CF مبتنی بر مدل نظیر SVD بر CF مبتنی بر حافظه در برخی موارد برتری دارد. بعلاوه، برخی پارامترهای CF می تواند اثرات مهمی بر صحت پیشگویانه مانند سنجش شباهت، اندازه مجاورت، و نسبت مجموعه اموزشی داشته باشد.

User ID	Video ID	Download	Play	Share	Like	Ratings
U1	A	√	√		√	3
U1	B	√				1
U2	A	√		√		2
U3	C	√				1

	A	B	C
U1	3	1	
U2		2	
U3			1

شکل ۱۰-ایجاد درجه بندی های تلویحی براساس ثبت های عملیاتی

A-مورد ۱: مبتنی بر CF روی رفتار کاربران

این مطالعه موردی با هدف توضیح این امر انجام شده که چگونه الگوریتم های CF را براساس رفتارهای کاربر اجرا نماییم. داده های دنیای واقعی از اپلیکیشن های موبایل یک پلتفورم ویدئویی با رشد سریع در چین جمع اوری شده است. بایگانی ثبت رفتارهای برخی کاربران جمع اوری شده و به پایگاه داده ها از طریق این اپلیکیشن ارسال گردید. بعد از خلاصی از داده های کثیف و حذف کاربرانی که ثبت عملیاتی شان کمتر از ۲۰ بوده است، تعداد ۱۱۳۱۰۵۳ ثبت در این مطالعه موردی رزرو شده که شامل ۱۶۰۸۲ مصرف کننده و ۱۹۸۲ ویدئو می باشد. در میان همه انواع رفتارهای کاربر، چهار تای آنها برای ارائه ترجیحات کاربر برای ویدئوها انتخاب شده است که عبارتند از دانلود، play ، Like و Share. الگوریتم های خاصی باید استفاده شود تا این ثبت های عملیاتی را به شکل

درجه بندی هایی تغییر شکل دهیم. در این مطالعه موردی، بخاطر سادگی کار، تعداد عملیات بالا که یک کاربر روی یک ویدئوی خاص اجرا کرده است به شکل درجه بندی های تلویحی انگاشته می شود که در شکل ۱۰ به تصویر درامده است. مشخص شده که درجه بندی تلویحی مبتنی بر رفتارهای کاربر از ۱ (بد) تا چهار (عالی) متغیر است و بازتاب علایق کاربر روی یک گزینه خاص می باشد. پایگاه داده های کامل به دو بخش تقسیم بندی شده است: ۸۰ درصد درجه بندی ها را به شکل مجموعه اموزشی انگاشته ایم و سایرین را به شکل مجموعه تست می دانیم. MAE تحت عنوان اصول سنجش صحت پیشگویی و شباهت میان کاربران انتخاب می شود یا گزینه ها با همبستگی پیرسون اندازه گیری می شود.

پیشنهادات تصادفی که درجه بندی مفقوده را به طور تصادفی پیشگویی می کند به شکل الگوریتم خط مبنا درنظر گرفته می شود. الگوریتم های پیشنهادات که در این مطالعه موردی درنظر گرفته شده است به ترتیب ذیل فهرست شده است:

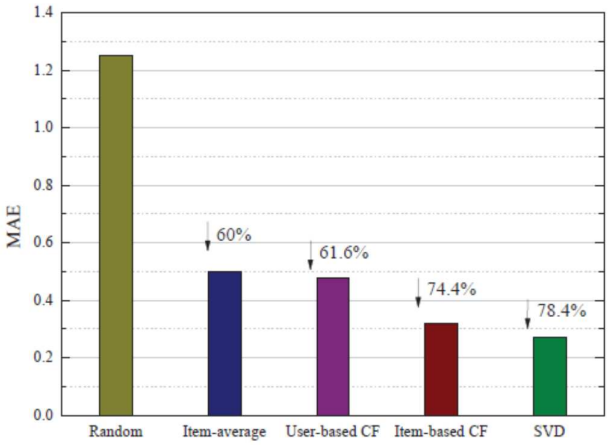
-تصادفی

-مبتنی برکاربر

-گزینه-متوسط

-مبنی بر گزینه

-SVD



شکل ۱۱-مقایسه MAE در میان الگوریتم های پیشنهادات معمول

طبق شکل ۱۱، صحت پیشگویی الگوریتم تصادفی پایین ترین است چون استفاده ای از داده ها نکرده است. CF مبتنی بر کاربر تا اندازه ای بهتر از الگوریتم متوسط گزینه طبق توجیه انتخاب مجاور صحیح تر می باشد. در این مطالعه موردی، CF مبتنی بر گزینه نسبت به CF مبتنی بر کاربر برتری دارد. یک علت احتمالی این است که تعداد کاربران خیلی فراتر از تعداد گزینه ها می باشد و باعث می شود که یافتن مجاوران مناسب مشکلتر بشود. SVD که نمایانگر CF مبتنی بر حالت می باشد دارای بالاترین صحت پیشگویی در این آزمایش است چرا که دو ماتریس ویژگی از طریق فاکتورگیری ماتریس برای توضیح ویژگی های کاربران و گزینه ها به طور اخص تر بدست آمده است. برخی تغییرات در نتایج در میان تغیرات برخی پارامترهای مهم وجود دارد. اثرات این پارامترها به طور مفصل در مطالعه موردی دوم تحلیل شده است.

B. مورد ۲: CF مبتنی بر درجه بندی های کاربر

این مطالعه موردی با هدف بحث درباره اثرات پارامترهای کلیدی روی صحت پیشگویانه انجام شده است که شامل اندازه گیری تشابه، اندازه مجاور و نسبت مجموعه اموزشی می باشد. پایگاه داده از Movielens برای اجرای این ازمایش انجام شده است. Movielens یک وب سایت تحقیقاتی است که توسط موسسه تحقیقاتی GroupLens Research در دانشگاه مینه سوتا راه اندازی شده است. صدها کاربر از این وب سایت دیدن کرده و فیلم ها را هر هفته درجه بندی کرده اند. موسسه تحقیقاتی GroupLens Research داده هایی را طی دوره هفت ماهه از ۱۹ سپتامبر ۱۹۹۷ تا ۲۲ اوریل ۱۹۹۸ جمع اوری کرده و پایگاه داده ها را با حذف کاربرانی که کمتر از ۲۰ فیلم را درجه بندی کرده بودند، فیلترسازی نمود. پایگاه داده های استفاده شده در این مطالعه موردی شامل صدهزار درجه بندی از ۹۴۳ کاربر درباره ۱۶۸۳ فیلم می باشد و درجه بندی ها از ۱ (بد) به ۵ (عالی) متغیر است. نیز، پایگاه داده به دو بخش مجموعه اموزشی و مجموعه تست تقسیم بندی می شود. متغیر الفا نمایانگر نسبت مجموعه اموزشی می باشد. برای مثال الفای برابر با ۰٫۸ نشان می دهد که ۸۰ درصد از داده ها به شکل مجموعه اموزشی انتخاب می شوند و بقیه ۲۰ درصد به شکل مجموعه تست انتخاب می شوند. هر دو الگوریتم های مبتنی بر کاربر و مبتنی بر گزینه اجرا شده و MAE برای اندازه گیری صحت پیشگویی الگوریتم های فوق انتخاب می شود.

عملکردهای سه سنجش ابتدا انالیز می شود از جمله کسینوس خالص، کسینوس تنظیم شده و PCC. سایر پارامترها به شکل ثابت تنظیم می شود: اندازه مجاور برابر با ۲۴ است و نسبت مجموعه اموزشی برابر با ۸۰ درصد می باشد. نتایج ازمایشی در شکل ۱۲a نشان داده شده است. می توان به وضوح مشاهده کرد که برای هر دو الگوریتم های مبتنی بر کاربر و مبتنی بر گزینه، PCC یک مزیت مشهودی دارد. بعلاوه، برای CF مبتنی بر کاربر کسینوس تنظیم شده همانند PCC می باشد که در بخش سوم نشان داده شده است.

اندازه مجاور نیز یک اثر مهم بر کیفیت پیشگویی دارد. از اینرو، یک ازمایشی برای محاسبه MAE انجام می شود حین اینکه اندازه مجاور از ۵ به ۸۰ با مقدار الفا که در ۰,۸ تثبیت شده است، متغیر است. نتایج ازمایشی در شکل ۱۲b و شکل ۱۲c نشان داده شده است. می توان به روشنی دید که همه این منحنی ها یک روند مشابهی دارد و اندازه مجاور بهینه هر سناریو به سهولت می تواند کسب شود. با درنظرگیری شرح برنامه CF مبتنی بر کاربر و PCC (خط سبز در شکل ۱۲b) به عنوان مثال، MAE به سرعت کاهش می یابد حین اینکه اندازه مجاور از ۵ به ۳۲ افزایش می یابد و بعد منحنی شروع به افزایش می کند. از اینرو، اندازه مجاور بهینه برابر با ۳۲ می باشد. این امر بدان خاطر است که در آغاز کار، رشد اندازه مجاور مستلزم مجاوران با تشابه بالا برای پیشگویی درجه بندی گزینه هدف می باشد. ولیکن، وقتی اندازه مجاور از مقدار بهینه تجاوز کرد، مجاوران با تشابه پایین درگیر می شوند که می تواند مجاورت را تنزل داده و منجر به افزایش MAE بشود. برای CF مبتنی بر کاربر، منحنی کسینوسی تنظیم شده روی منحنی PCC همپوشانی دارد، که در بخش سوم توضیحش داده شده است.

طبق ازمایشات فوق، CF مبتنی بر گزینه با PCC بهتر عمل می کند. از اینرو، این انتخاب انجام می شود که اثر الفا روی نتایج تحلیل شود. یک آزمایش اجرا می شود حین اینکه الفا از ۴۸ به ۸۰ درصد تغییر می نماید و فاصله مرحله برابر با ۴ درصد می باشد. اندازه مجاور نیز در این آزمایش در نظر گرفته شده است که از ۱۲ الی ۳۲ با فاصله گام برابر با ۴ متغیر است. نتایج ازمایشی در شکل ۱۲d نشان داده شده است. می توان مشاهده کرد که با افزایش الفا، صحت پیشگویی به یمن غنی بودن داده های آموزشی بهبود می یابد.

۵-نتیجه گیری

این مقاله درباره الگوریتم های CF بحث کرده که در اپلیکیشن های اینترنت موبایل بکار گرفته می شود. یک چارچوب کاری ابتدا برای نشان دادن روشهای اصلی یک سیستم پیشنهاددهنده CF معمولی یعنی جمع آوری

داده ها، پیش پردازش داده ها، و فیلترسازی مبتنی بر همکاری مطرح شده است. ویژگی های دو نوع داده های کاربر یعنی رفتارهای کاربر و درجه بندی کاربر تحلیل شده و به تفصیل مقایسه شده است. بعد از تغییر شکل رفتارهای کاربر به شکل درجه بندی های تلویحی از طریق روشهای خاص و ویژه، مسئله کمیابی ماتریس درجه بندی می تواند تا حدودی تخفیف داده شود. الگوریتم های CF معمولی شامل مبتنی بر حافظه و مبتنی بر مدل معرفی شده و عملیات عمومی آنها برای خاطر آشکارسازی ویژگی های معمول این روشها خلاصه بندی می شود. سرانجام اینکه، برای روایی سازی این چارچوب، دو مطالعه موردی براساس به ترتیب رفتارهای کاربر و درجه بندی های کاربر انجام گردید. هر مطالعه موردی برخی الگوریتم های CF معمولی را اجرا کرده است و بعد عملکردهای آنها را در صحت پیشگویانه با اصول سنجش MAE مقایسه کرده است. هرچند پیشرفت مهمی در مسئله CF اجرا شده است، مطالعات بیشتری همچنان در برخی جنبه ها مورد نیاز است. برای مثال، برخی سیستم های توصیه کننده در پردازش داده های جمعی در زمان محدود شده با قابلیت های ذخیره سازی و محاسباتی دچار کمبود می باشند. از اینرو، برای بهبود ظرفیت پردازش و کاهش مصرف زمانی، تدوین الگوریتم هایی برای سیستم های محاسباتی توزیع شده می تواند یک رهنمود تحقیقاتی برجسته در آینده باشد. باور بر این است که سیستم های پیشنهاددهنده CF که به طور مداوم بهبود می یابند می توانند تا حد زیادی به کاربران اینترنت موبایل کمک کنند تا گزینه های مناسبی را بدون مصرف زمان و انرژی اضافی در حیطه داده های بزرگ بیابند.

ترجمه فا
TarjomeFa.Com

این مقاله، از سری مقالات ترجمه شده رایگان سایت ترجمه فا میباشد که با فرمت PDF در اختیار شما عزیزان قرار گرفته است. در صورت تمایل میتوانید با کلیک بر روی دکمه های زیر از سایر مقالات نیز استفاده نمایید:

✔ لیست مقالات ترجمه شده

✔ لیست مقالات ترجمه شده رایگان

✔ لیست جدیدترین مقالات انگلیسی ISI

سایت ترجمه فا ؛ مرجع جدیدترین مقالات ترجمه شده از نشریات معتبر خارجی