
بسم الله الرحمن الرحيم

دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

ارائه روش جدید در سامانه‌های توصیه‌گر با رویکرد پالایش گروهی

کارشناسی ارشد مهندسی کامپیوتر- نرم‌افزار

امیرحسین قرائتی

اساتید راهنما

دکتر سید رسول موسوی
دکتر چنگیز اصلاح‌چی

۱۳۹۳

دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش نرم افزار

آقای امیرحسین قرائتی تحت عنوان

ارائه روش جدید در سامانه‌های توصیه‌گر با رویکرد پالایش گروهی

در تاریخ ۹۳/۱۰/۲۴ توسط کمیته تخصصی زیر مورد بررسی و تصویب نهایی قرار گرفت.

۱-استاد راهنمای اول پایان نامه	دکتر سید رسول موسوی
۲-استاد راهنمای دوم پایان نامه	دکتر چنگیز اصلاح‌چی
۳-استاد داور	دکتر عبدالرضا میرزائی
سرپرست تحصیلات تکمیلی دانشکده	دکتر محمدعلی خسروی فرد

فهرست

چکیده

سامانه‌های توصیه‌گر، ابزارهای نرم افزاری و تکنیک‌هایی هستند که برای کاربران پیشنهادهایی از آیتم‌ها تهیه می‌کنند. امروزه با رشد سریع وب و اطلاعات موجود بر روی آن، سامانه‌های توصیه‌گر توجه زیادی به خود جلب کرده‌اند. بسیاری از وب‌سایت‌های معتبر برای ارائه خدمات به کاربران خود از یک توصیه‌گر بهره می‌برند. به‌طور کلی، یک سامانه توصیه‌گر با رویکرد پالایش گروهی، از کاربران و آیتم‌ها تشکیل شده است، به شکلی که هر کاربر تعدادی از آیتم‌ها را در کتابخانه خود جمع آوری کرده و سامانه توصیه‌گر با توجه به این آیتم‌ها و آیتم‌هایی که دیگر کاربران جمع آوری کرده‌اند، آیتم‌های جدیدی که کاربر تا به حال جمع آوری نکرده و احتمالاً به آن‌ها علاقمند باشد، به او پیشنهاد می‌دهد.

تاکنون برای توسعه توصیه‌گرها از تکنیک‌ها و تکنولوژی‌های مختلفی استفاده شده است. ما در این تحقیق روش جدیدی در سامانه‌های توصیه‌گر با رویکرد پالایش گروهی ارائه می‌دهیم که بر اساس حافظه است و از تشابه بین دو کاربر برای تولید پیشنهاد استفاده می‌کند، همچنین با تعریف یک همسایگی وزن‌دار در بین آیتم‌ها، یک پس زمینه کلی از علاقه به آیتم‌ها را در نظر می‌گیریم که با عث افزایش دقت پیشگویی الگوریتم جدید می‌شود.

مجموعه داده ویژه‌ای با نام MovieLens برای ارزیابی توصیه‌گرها ایجاد شده، که اکثر توصیه‌گرها را با این مجموعه داده ارزیابی کرده‌اند. ما روش پیشنهادی در این پایان‌نامه را با روشی کاربر محور بر اساس همبستگی پیرسون که یکی از پرکاربردترین روش‌ها در بین روش‌های کاربرر محور است و روش‌های NBI، NBIw و INBI، که بر اساس دانش فعلی ما نتایج ارزیابی قابل قبولی در بین دیگر روش‌های آیتم محور دارند، مقایسه کرده‌ایم. نتایج آزمایشات با توجه به معیارهای ارزیابی، نشان می‌دهد روش پیشنهادی نتایج ارزیابی بهتری، نسبت به این روش‌ها دارد.

فصل اول
مقدمه‌ای بر سامانه‌های توصیه‌گر

۱-۱ معرفی سامانه توصیه‌گر

سامانه توصیه‌گر[۱](RS) یا سامانه پیشنهادگر، ابزارهای نرم‌افزاری و تکنیک‌هایی است که برای کاربر پیشنهادهایی از آیتم‌ها برای استفاده تهیه می‌کند. پیشنهادها مربوط به فرآیندهای مختلف تصمیم‌گیری هستند، مانند خریدن کدام کالا، گوش دادن به کدام موسیقی، یا خواندن کدام اخبار آنلاین. به‌طور کلی، از واژه آیتم برای معنی کردن آن‌چه سیستم توصیه می‌کند، استفاده می‌شود.

سامانه توصیه‌گر به‌طور مثال برای افرادی که تجربه شخصی و قدرت مقایسه کافی، برای انتخاب بین تعداد فراوان آیتم‌های موجود در یک وب سایت را ندارند، مفید است. برای مثال، سامانه توصیه‌گر در یک وب سایت فروش کتاب به کاربران کمک می‌کند کدام کتاب را برای خواندن انتخاب کنند. امروزه در سایت‌های پرطرفدار مانند Amazon.com از یک سامانه توصیه‌گر استفاده می‌کنند تا سایت را برای هر کاربر شخصی‌سازی[۲] کنند. سامانه‌های توصیه‌گر شخصی سازی نشده هم وجود دارند، مانند لیست پرفروش‌ترین کتاب‌ها. این توصیه‌گرها توسعه به مراتب

۱- Recommendation system
۲- Personalized

ساده‌تری دارند. با این‌که، این‌گونه توصیه‌گرها ممکن است در بعضی شرایط مفید باشند، ولی توصیه‌گرهای شخصی‌سازی نشده موضوع اصلی تحقیقات سامانه‌های توصیه‌گر نیستند.

در ساده‌ترین حالت، توصیه‌های شخصی‌سازی شده به شکل یک لیست امتیاز[۱] از آیتم‌ها هستند. در این لیست امتیاز سامانه توصیه‌گر تلاش می‌کند، مناسب‌ترین محصول یا سرویس را برای کاربر بر اساس ترجیحات او، پیشگویی کند. برای انجام چنین کاری، سامانه توصیه‌گر باید ترجیحات کاربران را که به طور معمول از امتیازی که کاربر به یک آیتم می‌دهد برداشت می‌شود، جمع‌آوری کند[۱].

توسعه سامانه‌های توصیه‌گر از یک ایده ساده شروع شد: اغلب افراد تصمیمات روزانه خود را بر اساس توصیه‌هایی که دیگران به آن‌ها می‌دهند، می‌گیرند. این توصیه‌ها می‌تواند به صورت کلامی باشد، و یا به شکل توصیه‌نامه، یا مقالات بررسی کتاب و فیلم که در روزنامه‌ها چاپ می‌شود. سامانه‌های توصیه‌گر سعی می‌کنند به این فرآیند طبیعی اجتماعی کمک کنند و آن را تقویت کنند[۲].

۱-۲ دلایل اهمیت سامانه‌های توصیه‌گر

در این سال‌ها علاقه به سامانه‌های توصیه‌گر افزایش چشم‌گیری داشته است، از جمله می‌توان به مواردی که در ادامه آورده می‌شود اشاره کرد.

رشد انفجاری اطلاعات موجود در وب و رشد سریع سرویس‌های تجارت الکترونیکی[۲] باعث ابهام کاربر در انتخاب بین گزینه‌های موجود می‌شود. سامانه توصیه‌گر نقش مهمی در سایت‌های معتبر اینترنتی از جمله Amazon.com، YouTube، Netflix و ... بازی می‌کند. برخی شرکت‌های رسانه‌ای سرمایه‌گذاری‌های کلانی بر روی توسعه سامانه‌های توصیه‌گر به‌عنوان بخشی از سرویسی که به مشتریان ارائه می‌دهند، می‌کنند. در حقیقت، افزایش فروش آیتم‌ها، دلیل اصلی استفاده از این تکنولوژی توسط ارائه دهنده‌های سرویس‌های مختلف بر روی اینترنت می‌باشد[۳].

از طرف دیگر، موارد زیادی تحقیق و توسعه در زمینه‌ی سامانه توصیه‌گر انجام می‌شود که در ژورنال‌های آکادمیک به چاپ می‌رسند؛ و همچنین برگزاری کنفرانس‌ها و کارگاه‌های آموزشی مربوط به این زمینه، بیانگر اهمیت موضوع سامانه توصیه‌گر بین محققان در این روزها می‌باشد[۱].

۱- Ranked list
۲- E-commerce

۱-۳ کاربردهای سامانه توصیه‌گر

در [۴] محققین یک رده‌بندی[۱] از سامانه‌های توصیه‌گر موجود در اینترنت ارائه دادند، سامانه‌های توصیه‌گر را از ابعاد گوناگون بررسی کرده و آن‌ها را دسته بندی کرده‌اند. با توجه به این رده‌بندی و بررسی سامانه‌های توصیه‌گر موجود بر روی اینترنت، می توان کاربردهای زیر را برای آن‌ها نام برد :

- تجارت الکترونیک : به کاربران آیتم‌هایی برای خرید بر روی یک فروشگاه آنلاین پیشنهاد می‌دهند مانند Amazon.com.
- سرگرمی : پیشنهاد فیلم و موسیقی به کاربران مطابق سلیقه آن‌ها مانند Netflix و IMDB.
- شبکه‌های اجتماعی : پیشنهاد افراد جدید برای ارتباط، مبتنی بر دوستان مشترک مانند Facebook و....

۱-٤ انواع سامانه‌های توصیه‌گر

سامانه‌های توصیه‌گر به‌طور کلی به سه دسته تقسیم می‌شوند؛ که شامل : ۱-پالایش گروهی[۲](CF)، ۲-محتوا محور[۳]، و گونه سومی تحت عنوان سامانه توصیه‌گر ترکیبی[۴] است.

۱-٤-۱ رویکرد پالایش گروهی

استفاده از الگوریتم‌های پالایش گروهی، یک رویکرد به سامانه‌های توصیه‌گر است که استفاده‌های بسیار زیادی دارد. رویکرد پالایش گروهی براساس جمع‌آوری و بررسی مقدار زیادی از اطلاعات رفتارها، فعالیت‌ها یا ترجیحات کاربران است، تا با توجه به شباهت کاربران، آن‌چه کاربران انتخاب خواهندکرد را پیشگویی کند. از فواید اصلی رویکرد پالایش گروهی به این نکته می‌توان اشاره کرد که نیاز به محتویات پیچیده و قابل فهم برای ماشین نیست، در نتیجه این قابلیت را دارد تا آیتم‌های پیچیده مانند فیلم‌ها را، بدون احتیاج به فهمیدن خود آیتم، به شکل دقیقی پیشنهاد دهد[۵].

در این رویکرد فرض بر این است که اگر کاربر A دارای سلیقه‌ی مشابهی با کاربر B باشد، در مقایسه با یک کاربر تصادفی، در یک موضوع مشترک x، دارای عقیده مشابهی با B خواهد بود. به‌طور کلی، پالایش گروهی بر اساس این

۱- Taxonomy
۲- Collaborative Filtering
۳- Content Base
٤- Hybrid

فرض است که افرادی که در گذشته با هم هم‌نظر بوده‌اند در آینده هم خواهند بود و اگر در گذشته آیتم‌های مشابهی را دوست داشته‌اند، به احتمال زیاد در آینده نیز آیتم‌های مشابهی را دوست خواهند داشت[۶].

در رویکرد پالایش گروهی جمع‌آوری اطلاعات به دو شکل صریح[۱] یا ضمنی[۲] انجام می‌شود. در شکل صریح کاربر تمایلات خود را در مورد یک آیتم به طور مستقیم بیان می‌کند، معمولاً به شکل امتیاز عددی بین ۱ تا ۵. در شکل ضمنی رفتار و یا انتخاب‌های کاربر برای بدست آوردن تمایلات او ترجمه می‌شوند، برای مثال سابقه خرید کاربر در یک سایت فروش آنلاین.

این رویکرد یک مدل از تمایلات کاربر فعال[۳] وتمایلات دیگر کاربران می‌سازد (با اطلاعاتی که به شکل صریح یا ضمنی جمع‌آوری می‌کند). سپس، از مقایسه مدل‌ها، آیتم‌هایی که کاربر ممکن است به آنها علاقه داشته باشد (یا امتیازی که به آن‌ها می‌دهد) را پیشگویی می‌کند. در این‌جا منظور از کاربر فعال، کاربری است که می‌خواهیم به او پیشنهاد بدهیم.

Amazon.com، Facebook، LinkedIn و ... مثال‌های تجاری و غیر تجاری هستند که از سامانه توصیه‌گری با رویکرد پالایش گروهی استفاده می‌کنند.

۱-٤-۲ مشکلات رویکرد پالایش گروهی

سامانه‌های توصیه‌گر با رویکرد پالایش گروهی اغلب با مشکلاتی همچون شروع سرد[۴]، مقیاس‌پذیری[۵] و پراکندگی[۶] مواجه هستند.

مشکل شروع سرد وقتی پیش می‌آید که سامانه توصیه‌گر بخواهد برای کاربری که تاکنون هیچ آیتمی را انتخاب نکرده است، پیشنهاد تولید کند. در چنین شرایطی از آن‌جایی که هیچ اطلاعاتی راجع به تمایلات کاربر در اختیار نداریم رویکرد پالایش گروهی به تنهایی از پس این مشکل بر نمی‌آید. در عمل از ترکیب این رویکرد با رویکرد محتوا محور برای حل این مشکل استفاده می‌شود[۷].

۱- Explicit
۲- Implicit
۳- Active User
٤- Cold Start
٥- Scalability
٦- Sparsity

از آن‌جایی که این سیستم‌ها، اکثراً در محیط‌هایی که میلیون‌ها کاربر و محصول وجود دارد تولید پیشنهاد می‌کنند، به قدرت محاسباتی زیادی برای بررسی سوابق کاربران نیاز دارند[۸]. به‌طور مثال، در سایت Twitter میلیون‌ها کاربر وجود دارد، ولی با به‌کارگیری تکنیک خاصی قدرت محاسباتی مورد نیاز سامانه توصیه‌گر را کاهش می‌دهند، تا سیستم بتواند پیشنهادهای آنلاین برای کاربران تولید کند[۹].

مشکل پراکندگی به این مسئله بر می‌گردد که با وجود میلیون‌ها آیتم، یک فرد به تنهایی نمی‌تواند حتی یک درصد از کل آیتم‌ها را بررسی کند[۸]. دربسیاری از موارد یک آیتم تنها توسط یک کاربر انتخاب شده ویا یک کاربر تنها یک آیتم را انتخاب کرده است. به‌طور مثال، در سایت Amazon.com چند میلیون آیتم وجود دارد که از یک کاربر انتظار نمی‌رود بیشتر از ۱۰،۰۰۰ تا از این آیتم‌ها را بررسی کرده باشد. مشکل پراکندگی، پیشنهاد به کاربر را برای توصیه‌گر دشوار می‌کند.

۱-٤-۳ رویکرد محتوا محور

از آنجایی که تولید پیشنهاد با رویکرد پالایش گروهی سوابق رفتاری کاربران را بررسی می‌کند، به قدرت محاسباتی بالایی برای تولید پیشنهاد احتیاج دارد، ویا ممکن است بحث حریم خصوصی کاربران پیش بیاید. ولی رویکرد محتوا محور، بدون نیاز به تطابق تمایلات کاربران و تنها بر اساس اطلاعاتی که خود آیتم به همراه دارد، می‌تواند پیشنهاد تولید کند[۱۰].

در سیستم‌های توصیه‌گر محتوا محور، از واژه‌های کلیدی برای توصیف آیتم‌ها استفاده می‌شود، به‌علاوه، یک پروفایل برای کاربر ساخته می‌شود تا انواع آیتم‌هایی که او دوست دارد در آن ذخیره شود. به بیان دیگر، این الگوریتم‌ها تلاش می‌کنند آیتم‌هایی را پیشنهاد بدهند که به آیتم‌هایی که کاربر در گذشته انتخاب کرده شباهت دارند. در عمل، چندین آیتم کاندید با آیتم‌هایی که کاربر در گذشته امتیاز داده مقایسه می‌شوند و آیتم‌هایی که بهترین هم‌خوانی را داشته باشند پیشنهاد می‌شوند[۱۱].

برای مثال، با دانستن ژانر فیلم‌ها و این‌که کاربر به چه ژانری از سینما علاقه دارد، می‌توانیم به کاربر فیلم‌های جدیدی پیشنهاد بدهیم. در کل، سامانه توصیه‌گر محتوا محور به روشی گفته می‌شود، که با مقایسه محتویاتی که یک آیتم را توصیف می‌کند، با محتویات مورد علاقه کاربر، پیشنهاد تولید می‌کند. برای توسعه این رویکرد، یا از روش بازیابی

اطلاعات استفاده می‌شود و یا از تکنیک‌های یادگیری ماشین از جمله کلاسبند بیزین[۱]، درخت تصمیم[۲] و شبکه‌های عصبی[۳][۱۲].

Pandora یک مثال بسیار خوب برای سامانه توصیه‌گر با رویکرد محتوا محور است[۱۳]. Pandora یک ایستگاه رادیویی آنلاین است و برای توصیف آهنگ ها از ۴۰۰ صفت متفاوت استفاده می‌کند. پایگاه داده‌های Pandora را پروژه ژنوم موسیقی[۴] می‌نامند. روزانه افراد زیادی در این کمپانی به بررسی آهنگ‌های جدید و مقداردهی به این صفات مشغول هستند تا سیستم بتواند برای کاربران پیشنهادهای بهتری تولید کند.

۱-٤-٤ رویکرد ترکیبی

گونه سوم توصیه‌گرها، سیستم‌های ترکیبی هستند که از ترکیبی از دو نوع رویکرد گفته شده استفاده می‌کنند. طراحان، این نوع سیستم‌ها را غالباً به دو منظور با هم ترکیب می‌کنند : ۱-افزایش عملکرد سیستم، ۲- کاهش اثر نقاط ضعف رویکردها وقتی به تنهایی به‌کار گرفته می‌شوند. تحقیقات اخیر نشان داده که ترکیب رویکردهای پالایش گروهی و محتوا محور می‌تواند در بعضی شرایط کاراتر باشد. رویکرد ترکیبی می‌تواند از راه‌های مختلفی پیاده‌سازی شود، می‌توان پیشبینی‌های رویکردهای پالایش گروهی و محتوا محور را جداگانه محاسبه کرد، وسپس آن‌ها را باهم ترکیب کرد، می‌توان توانایی‌های رویکرد محتوا محور را به رویکرد پالایش گروهی اضافه کرد و بالعکس، یا می‌توان رویکردها را در یک مدل یکی کرد. در چندین مطالعه مشاهداتی، کارایی رویکرد ترکیبی را با رویکردهای پالایش گروهی و محتوا محور به تنهایی، مقایسه کرده‌اند و نشان داده‌اند که رویکرد ترکیبی می‌تواند پیشنهادهای دقیق‌تری نسبت به دو رویکرد مذکور به تنهایی، ایجاد کند.همچنین این رویکرد می‌تواند برای غلبه بر مشکلاتی همچون شروع سرد و یا پراکندگی استفاده شود[۱۴].

Netflix یک مثال خوب از رویکرد ترکیبی است[۶]. آن‌ها، هم با مقایسه عادت‌های کاربران در مشاهده فیلم، به کاربرها فیلم پیشنهاد می‌دهند، و همین‌طور فیلم‌هایی را پیشنهاد می‌دهند که مشخصه‌های مشابه با فیلم‌هایی که کاربر به آن‌ها امتیاز بالا داده را دارا است.

۱- Baysian Classifier
۲- Dicision Tree
۳- Artificial Neural Network
٤- Music Genome Projectی

۱-٤-٥ مقایسه رویکردهای پالایش گروهی و محتوا محور

اولین تلاش‌ها برای پالایش اطلاعات، بر اساس محتوا بود. این سیستم‌ها، آیتم‌ها را براساس محتوای آن‌ها برای پیشنهاد انتخاب می‌کنند. بنابراین، پروفایل کاربر بیان کننده محتواهایی است که کاربر به آن‌ها علاقه دارد. این گونه از پالایش اطلاعات وقتی کارا است که داده های متنی را بازیابی می‌کنیم، به طوری که هر سند[۱] از مجموعه‌ای از کلمات کلیدی تشکیل شده باشد. به هر حال این سیستم ها محدودیت‌هایی دارند[۱۵].

اول از همه این‌که آیتم‌ها باید توسط ماشین قابل بررسی باشند. وقتی اطلاعات چند رسانه‌ای را بازیابی می‌کنیم، در شرایطی که ادراک ماشین از محتوا نسبت به ادراک کاربر از محتوا خیلی ضعیف‌تر است، این کار دشوار می‌شود. اگرچه انتساب صفات توسط یک فرد این مشکل را حداقل در این زمینه حل می‌کند، ولی رویکرد محتوا محور در برخورد با بسیاری از اطلاعات امروزی نامناسب است[۱۵].

مشکل دیگر پالایش محتوا محور عدم توانایی در ارزیابی کیفیت یک آیتم است. برای مثال، این روش نمی‌تواند تمایزی بین یک مقاله خوب و یک مقاله بد قایل شود اگر هر دو مقاله از لغات یکسانی استفاده کرده باشند. در حقیقت، کیفیت یک آیتم یک ویژگی خیلی موضوعی است که بستگی به سلایق، ایده‌ها، فرهنگ و ... از هر شخص دارد و ممکن است برای ماشین دشوار باشد تا متوجه شود[۱۵].

در نهایت، پالایش محتوا محور راهی ندارد تا آیتم‌های غیر مترقبه[۲] که برای کاربر جالب هستند را پیدا کند، این‌ها آیتم‌هایی هستند که به طور معمول به پروفایل کاربر مربوط نیستند[۱۵].

سیستم‌های پالایش گروهی کمتر به این مشکلات حساس هستند، از آن‌جایی که به‌جای محتوای آیتم‌ها براساس سلایق کاربران است. این سیستم، آیتم‌هایی را پیشنهاد می‌کند که از کاربرانی که سلایق مشابهی دارند، امتیاز بالایی دریافت کرده باشند. در این تکنیک‌ها، آیتم‌ها توسط افراد امتیازدهی می‌شوند و سیستم نیازی ندارد که محتوای آیتم‌ها را بررسی کند و کیفیت یا ارزیابی موضوعی یک آیتم هم در نظر گرفته شده است[۱۵].

۱- document
۲- serendipitous

رویکرد پالایش گروهی سه فایده کلیدی برای پالایش اطلاعات[۱] فراهم می‌کند که هیچ یک از آن‌ها توسط رویکرد محتوا محور فراهم نشده است: (۱) پشتیبانی از پالایش آیتم‌هایی که پردازش محتوای آن‌ها برای یک فرآیند خودکار ساده نیست. (۲) توانایی پالایش آیتم‌ها بر اساس کیفیت[۲] و سلیقه[۳]. (۳) توانایی تولید پیشنهادهای غیر مترقبه[۱۶].

اول از همه، در پالایش گروهی ارتباط، کیفیت و علاقه به یک آیتم در اطلاعات، توسط انسان‌ها مشخص شده است. به‌عنوان مثال، می‌توانیم آیتم‌هایی را پالایش کنیم که آنالیز آن‌ها برای کامپیوترها دشوار است، مانند فیلم‌ها، ایده‌ها، احساسات، مردم و سیاستمدارن[۱۶].

دوم، پالایش گروهی می‌تواند در ابعادی فراتر از یک محتوای ساده، اندازه‌گیری کند که یک آیتم به چه خوبی نیاز یا علاقه کاربر را برآورده می‌کند که باعث بهبود پالایش اطلاعات می‌شود. انسان‌ها قادر هستند آیتم‌ها را در ابعادی مانند کیفیت یا سلیقه بررسی کنند که پردازش آن‌ها برای کامپیوترها خیلی سخت است. یک جستجوگر محتوا محور نشریات می‌تواند تمام مقالات راجع به یک اتفاق را بازگرداند، ولی تنها با ترکیب با یک رویکر پالایش گروهی می‌تواند مقالات مرتبط به آن موضوع که خوب نوشته شده‌اند را، بازگرداند[۱۶].

در نهایت ، یک رویکرد پالایش گروهی در بعضی مواقع پیشنهادهای غیر مترقبه تولید می‌کند، آیتم‌هایی را پیشنهاد می‌دهد که برای کاربر خیلی ارزشمند هستند، اما شامل محتوایی که کاربر انتظار آن را داشته است، نیستند. ما متوجه شدیم که پیشنهادهای غیر مترقبه معمولاً در دامنه فیلم اتفاق می‌افتد، با سیستم پالایش گروهی فیلم‌هایی را به‌طور دقیق پیشنهاد می‌دهیم که در شرایط دیگری کاربر آن‌ها را بررسی نمی‌کرد[۱۶].

۱-۵ اهداف این پایان‌نامه

در ابتدا هدف ما انجام تحقیقی در زمینه بیوانفورماتیک و موضوع پیشگویی تعاملات بین پروتئین‌ها و داروها[۴] بود. وقتی می‌گوییم پروتئین و دارو با هم تعامل دارند یعنی این دو ساختاری دارند که به هم متصل می‌شوند، ولی از آن‌جایی که ساختار تمام پروتئین‌ها شناخته شده نیست، اکثر تعاملات بین پروتئین‌ها و داروها نیز ناشناخته مانده است. در تحقیقات مربوط به پیشگویی تعاملات بین پروتئین و دارو هدف پیشگویی تعاملات ناشناخته بر اساس تعاملات شناخته شده است تا محدوده جستجو تعاملات جدید را برای محققان محدود کند. در طول تحقیق به مقاله‌ای برخوردیم که در آن برای پیشگویی تعامل بین پروتئین و دارو از یک سامانه توصیه‌گر که از روش NBI بهره می‌برد، استفاده کرده بود[۱۷]. در آن مقاله، تعاملات شناخته شده بین پروتئین‌ها و داروها را همانند ارتباط‌های کاربران و فیلم‌ها، با یک گراف دوبخشی

۱- Information filtering
۲- quality
۳- taste
۴- Drug-target interaction prediction

مدل می‌کرد و به‌عنوان ورودی به یک توصیه گر NBI می‌داد و در نهایت پیشنهادهای توصیه‌گر که در واقع پیشگویی تعاملات جدید ناشناخته بین پروتئین و دارو بودند را به‌عنوان خروجی می‌داد. ما به این نتیجه رسیدیم که اگر بتوانیم توصیه‌گری با دقت پیشگویی بهتر از NBI توسعه دهیم، قادر خواهیم بود پیشگویی دقیق‌تری از تعاملات بین داروها و پروتئین‌ها داشته باشیم.

با توجه به ماهیت مسئله پیشگویی تعاملات بین پروتئین‌ها و داروها باید از توصیه‌گری با رویکرد پالایش گروهی استفاده شود، زیرا اطلاعات کاملی راجع به محتوای پروتئین‌ها یا همان ساختار آن‌ها نداریم و تنها اطلاعاتی راجع به برخی تعاملات آن‌ها با داروهای مختلف را داریم. همان‌طور که گفتیم رویکرد پالایش گروهی مشکلات اساسی خود را دارد ولی در کاربردی که مد نظر ما بود این مسایل خیلی مشکل ساز نبودند و مسئله اصلی تولید پیشنهاد با دقت بیشتر بود. در نتیجه هدف ما به ارائه روش جدیدی در سامانه‌های توصیه‌گر با رویکرد پالایش گروهی که دقت پیشگویی بهتری از روش‌های موجود شناخته شده داشته باشد، تغییر کرد.

۱-٦ نوآوری‌ها در این پایان‌نامه

در این تحقیق روش جدیدی در سامانه‌های توصیه‌گر با رویکرد پالایش گروهی ارائه شده که در فصل سوم بیشتر توضیح داده خواهد شد. در این روش که بر اساس حافظه است ما از تشابه بین کاربران استفاده می‌کنیم. یکی از نوآوری‌های این تحقیق این است که برای اندازه‌گیری تشابه میان دو کاربر فرمول جدیدی ارائه می‌کنیم که کاستی‌های فرمول‌هایی که در روش‌های قدیمی استفاده می‌شود را پوشش می‌دهد. در نوآوری دیگری نسبت به روش های قدیمی بر اساس تشابه بین دو کاربر، با استفاده از یک همسایگی وزن‌دار در بین آیتم‌ها، دقت پیشگویی الگوریتم خود را افزایش می‌دهیم. این همسایگی وزن‌دار بیان‌گر پس زمینه‌ای کلی از علاقه به آیتم‌ها است

در ادامه نشان می‌دهیم که معیارهای ارزیابی موجود برای بیان دقیق کارایی الگوریتم‌ها کافی نیست و ما معیار دقت جدیدی تعریف می‌کنیم که بیان دقیق‌تری از کارایی روش‌ها دارد.این معیار دقت جدید بر اساس جایگاه پیشنهادها در لیست پیشنهاد است و به طول لیست پیشنهاد بستگی دارد. این معیار جدید در کنار معیار نرخ برخورد می‌تواند بیان دقیقی از کارایی الگوریتم در یک پیشنهاد با طول مشخص داشته باشد.

۱-۷ مروری بر فصل‌های این پایان‌نامه

در این پایان‌نامه، ابتدا در فصل دوم، روش‌های توصیه‌گرهای موجود با رویکرد پالایش گروهی را بررسی می‌کنیم و انواع آن‌ها را توضیح می‌دهیم. از آن‌جایی که روش جدید بر اساس حافظه است، تمرکز ما در این فصل نیز بیشتر بر روی روش‌های بر اساس حافظه خواهد بود. از بین این روش‌ها، روش‌هایی که از هبستگی پیرسون برای اندازه‌گیری

تشابه بین دو کاربر استفاده می‌کنند و کاربر محور هستند را به طور کامل توضیح می‌دهیم که یکی از پرکاربردترین روش‌ها از نوع کاربر محور است و سپس روش‌های استنتاج بر مبنای گراف دوبخشی و سیر تکامل آن‌ها را مفصل توضیح می‌دهیم که در بین روش‌های موجود بر اساس دانش فعلی ما نتایج ارزیابی قابل قبولی را دارند.

پس از آن در فصل سوم، روش جدید را ارائه می‌دهیم. در این فصل ابتدا معیار جدیدی برای تشابه بین دو کاربر تعریف می‌کنیم که کاستی‌های معیارهای پیشین را پوشش می‌دهد، سپس با تعریف همسایگی وزن دار در بین آیتم‌ها یک پس زمینه کلی از علاقه به آیتم‌ها در نظر می‌گیرم و دقت الگوریتم جدید را در تولید پیشنهاد افزایش می‌دهیم.

در فصل چهارم نتایج ارزیابی روش جدید را با نتایج روش‌های توضیح داده شده در فصل دوم مقایسه می‌کنیم. در این فصل با استفاده از مجموعه داده معروفی که برای این کار ایجاد شده است و معیارهای ارزیابی موجود و معیار دقت جدیدی که تعریف می‌کنیم، روش جدید را در کنار روش‌های نام برده ارزیابی می‌کنیم و نشان می‌دهیم که روش جدید نتایج ارزیابی بهتری نسبت به روش‌های نام برده شده دارد.

و در نهایت در فصل پنجم این پایان‌نامه، ابتدا جمع‌بندی از آن‌چه گفتیم می‌کنیم و سپس پیشنهادهایی برای بهبود روش جدید ارائه می‌دهیم همچنین کاربردهای الگوریتم جدید در کارهای آتی را بیان می کنیم. در آخر چالش‌های باقی‌مانده در زمینه سامانه‌های توصیه‌گر را نام می‌بریم.

فصل دوم

بررسی روش‌هایی که با رویکرد پالایش گروهی ارئه شده‌اند

۲-۱ مقدمه

همان‌طور که در فصل پیش گفتیم، پالایش گروهی عادات هر کاربر در استفاده از اطلاعات مختلف را بررسی می‌کند. رویکرد پالایش گروهی فواید بیشتری از رویکر محتوا محور دارد، مانند سادگی در نصب و توانایی پالایش هرگونه اطلاعات یا کالا بدون در نظر گرفتن محتوای آن‌ها. نتایج تجربی نیز نشان داده‌اند که رویکرد پالایش گروهی در بسیاری از مواقع از رویکرد محتوا محور عملکرد بهتری دارد[۱۸]. در حال حاظر پالایش گروهی موفق‌ترین و پراستفاده‌ترین رویکرد در سامانه‌های توصیه‌گر است[۱۴]. تمرکز ما در این تحقیق نیز بر روی روش‌های رویکرد پالایش گروهی است.

در سیستم‌های بر اساس پالایش گروهی، پروفایل کاربر شامل مجموعه‌ای از امتیازات داده شده به آیتم‌ها است. این امتیازات با پرسش از کاربر می‌توانند به شکل مستقیم دریافت شوند یا به شکل ضمنی با مشاهده تعامل کاربر با سیستم. امتیازدهی معمولاً به شکل یک ارزش یکتا است(آیتم‌های مربوط را نشان می‌دهد) یا یک مقدار دودویی(اجازه می‌دهد بین آیتم‌های خوب و بد تمایز قایل شویم) یا به طور معمول از یک ارزش عددی در یک مقیاس ثابت است[۱۵].

امتیازات کاربر در یک جدول که به نام ماتریس امتیاز شناخته می‌شود ذخیره می‌شود. این جدول برای تولید پیشنهاد پردازش می‌شود. بسته به این که داده‌های ماتریس امتیاز چگونه پردازش شوند دو نوع الگوریتم متفاوت داریم: براساس حافظه و براساس مدل[۱۵].

الگوریتم‌های براساس حافظه از تمام جدول برای محاسبه پیشگویی خود استفاده می‌کنند. به‌طور کلی، آن‌ها از معیارهای تشابه استفاده می‌کنند تا کاربرانی که به کاربر فعال شبیه هستند را انتخاب کنند. سپس، پیشگویی از امتیازاتی که همسایه‌ها داده‌اند محاسبه می‌شود(به همین علت به آن‌ها بر اساس همسایه نیز گفته می‌شود). اکثر این الگوریتم‌ها می‌توانند براساس این که تمرکز بر روی پیدا کردن کاربران متشابه هست یا آیتم های متشابه به دو دسته کاربر محور و آیتم محور تقسیم شوند[۱۵].

الگوریتم‌های براساس مدل ابتدا یک مدل می‌سازند که نشان دهنده رفتار کاربر است و سپس امتیازدهی کاربران را پیشگویی می‌کنند. پارامترهای مدل به شکل آفلاین از ماتریس امتیاز برآورده می‌شود. در مقالات رویکردهای متفاوتی برای این روش وجود دارد، بیشتر آن‌ها مربوط به یادگیری ماشین، ماشین بردار پشتیبان(SVM)[۱][۱۹]، بر اساس روش جبر خطی، بررسی فاکتور، خوشه بندی، شبکه‌های عصبی، گراف‌ها یا روش احتمالاتی مانند شبکه‌های بیزین است[۲۰].

به‌طور کلی الگوریتم‌های براساس حافظه ساده‌تر هستند، در حالی که نتایج با دقت قابل قبولی دارند. به هر حال، از آن‌جایی که باید تمام داده را برای یک پیشنهاد پردازش کنند، آن‌ها مشکلات جدی پراکندگی دارند. با تعداد زیادی از کاربران و آیتم‌ها این الگوریتم‌ها برای سیستم‌های آنلاین که آیتم‌ها را در زمان واقعی پیشنهاد می‌دهند مناسب نیستند. همچنین، آن‌ها بیشتر از روش‌های بر اساس مدل به مشکلات اساسی سامانه‌های توصیه‌گر حساس هستند. این مشکلات توسط الگوریتم‌های بر اساس مدل کاهش یافته است[۱۵].

۲-۲ الگوریتم‌های کاربر محور و آیتم محور

الگوریتم‌های کاربر محور، با نام همسایه محور نیز شناخته می‌شوند و یکی از محبوب‌ترین استراتژی‌ها در بین روش‌های پالایش گروهی است. این الگوریتم‌ها از یک فرآیند سه مرحله ای تشکیل شده‌اند[۱۵]:

۱) تشابه بین کاربر فعال و دیگر کاربران را محاسبه می‌کنند.

۲) یک زیرمجموعه از کاربران(همسایه‌ها) براساس شباهت آن‌ها با کاربر فعال پیدا می‌کنند.

۱- Support Vector Machine