تحلیل الگوریتمهای پالایش مشارکتی مبتنی بر کالا در ارایه خدمات به شهروند الکترونیک

محمددرزی[1]،حبیب اله اصغری[2]،آندره شلتز(Szelc Andrzej Stanislaw)[3]
[1]عضو شورای علمی گروه ITBMپژوهشکده ICTجهاددانشگاهی
Modarzi@yahoo.com
[2]عضو هیات علمی و رییس مرکز رشد فناوری اطلاعات و ارتباطات جهاددانشگاهی
Asghari@itincubator.com
[3]عضو هیات علمی ومعاون بین الملل دانشگاه UITM
Aszelc@wsiz.rzeszow.pl

۱- مقدمه:

حجم اطلاعات در دنیای امروز بسیارسریعتر از قدرت و توان ما در پردازش اطلاعات، روبه افزایش است. همه ما هر ساله با تولید تعداد جدیدی کتاب، مقالات و نشریات و ویژه‌نامه‌های کنفرانسی، افزایش روزافزون اطلاعات را احساس می‌کنیم. تکنولوژی، محدودیتهای چاپ و توزیع اطلاعات را کاهش داده است. هم اکنون زمان آن رسیده که تکنولوژیهایی ایجاد شوند که بتوانند درغربال کردن اطلاعات موجود برای پیدا کردن آنچه که برای ما با ارزش تر است، به ما کمک کنند.

یکی ازاین تکنولوژی‌ها، سیستمهای پیشنهاددهنده است. تکنولوژی‌های متعددی دراین سیستمها مورد استفاده قرار می‌گیرد که الگوریتم های پالایش مشارکتی یکی از آنهاست[14,16,19,27].الگوریتم پالایش مشارکتی با تشکیل پایگاه داده اولویتها برای کالاها توسط کاربران، اجرا می شود.این تکنولوژی هم درحوزه تحقیق و هم در پیاده سازی وهم درکاربردهای پالایش اطلاعات و هم تجارت الکترونیکی بسیار موفق بوده است. هرچند سوالات تحقیقی مهمی در مقابله با دو چالش اساسی سیستم های پیشنهاد دهنده پالایش مشارکتی باقیمانده است.

بحث و سوال اول، بهبود مقیاس‌پذیری الگوریتمهای پالایش مشارکتی است. این الگوریتمها قادرند در میان دهها هزار همسایه احتمالی بلادرنگ جستجو کنند، اما سیستم های مدرن، نیازمند جستجو در میان دهها میلیون از همسایه‌های احتمالی هستند. علاوه براین، الگوریتمهای موجود مورد استفاده در سایتها که حجم زیادی از اطلاعات را در اختیار دارند، مشکلات اجرایی در مورد کاربران خود دارند.

بحث دوم، در مورد بهبود کیفیت پیشنهادات برای کاربران است. کاربران به پیشنهاداتی نیاز دارند که بتوانند برای کمک در پیدا کردن کالاهایی که دوست دارند، به آنها اعتماد کنند. کاربران با رد استفاده ازسیستم های پیشنهاد دهنده که پیشنهادات آن متناسب با سلایق آنها نیست،عملا بی اعتنایی خود را به اینگونه سیستم‌ها نشان می‌دهند.

از برخی زوایا، این دو بحث با هم درتعارض است، به طوری که الگوریتمی که زمان کمی برای جستجوی همسایه ها صرف می کند، مقیاس پذیرتر خواهد بود ولی کیفیتش پایین تر . به همین علت، مهم است که این دو مساله به طور همزمان مورد بحث قرار گیرند و راه حلهای پیدا شده برای هردو، مفید و عملی باشد.

در این مقاله، این مشکلات، ازطریق به کار گیری الگوریتم های مبتنی بر کالا[1] بررسی می شود. معضل متداول در الگوریتمهای پالایش مشارکتی، جستجو برای همسایه‌ها در بین جمعیت زیاد کاربران همسایه است[12]. الگوریتمهای مبتنی بر کالا از این معضل جلوگیری می کنند، به این صورت که ابتدا ارتباط بین کالاها بدست می آیدتا ارتباط بین کاربران.

پیشنهادات برای کاربران از طریق پیدا کردن کالاهایی که مشابه دیگر کالاهایی هستند که کاربر به آنها علاقه داشته، محاسبه می شود. چون ارتباطات بین کالاها نسبتاً ایستا است، الگوریتمهای مبتنی بر کالا می توانند کیفیت مشابهی با الگوریتم های مبتنی بر کاربر [2] ولی با محاسبات برخط [3] کمتر، تولید کنند.

۱-۱- تاریخچه:

در این قسمت،ما پیشینه کوتاهی از ادبیات تحقیقی مرتبط با پالایش مشارکتی،سیستمهای پیشنهاددهنده، داده کاوی و شخصی‌سازی را ارایه می کنیم.

Tapestry [10] یکی از قدیمی‌ترین سیستم های پیشنهاددهنده است که مبتنی بر پالایش مشارکتی بود.این سیستم مبتنی بر ایده های آشکار افراد یک گروه مانند یک گروه کاری در یک اداره بود.به هر حال یک سیستم پیشنهاددهنده برای یک جامعه بزرگ نمی تواند وابسته به فرد یا افرادی باشد که آنها دیگران را بشناسند. پس از این سیستمهای پیشنهاددهنده خودکارمتعددی توسعه یافتند. سیستم پژوهشی GroupLens [16,19] یک راه حلی را برای

فیلمها و اخبار یوس نت [۴] از طریق پالایش اطلاعات ارایه کرد. Ringo [27] و پیشنهاددهنده Video [14] سیستمهای مبتنی بر پست الکترونیک و وب هستند که به ترتیب برای موزیک و فیلم پیشنهاد ارایه می کنند.

همچنین تکنولوژیهای دیگری مانند شبکه های بیزین [۵]، خوشه بندی[۶] و گراف هرتینگ[۷] در سیستمهای پیشنهاددهنده به کار گرفته شده است. شبکه های بیزین یک مدل را بر مبنای یک مجموعه آموزشی [۸] با یک درخت تصمیم ایجاد می کند که هر گره و یا ل بیان کننده اطلاعات کاربر است.این مدل می تواند به صورت غیر بر خط[۹] ساخته شود.نتیجه مدل بسیار کوچک ، سریع و ضرورتا به اندازه روشهای نزدیک ترین همسایه [۱۰] دقیق است[6].

تکنیک های خوشه بندی با تعیین کردن گروههای کاربران که مشخص می شود سلایق مشترک دارند،عمل می کند.از زمانی که خوشه ها ایجاد می شوند،پیش بینی ها برای یک فرد می تواند از طریق گرفتن میانگین ایده های دیگر کاربران حاضر در آن خوشه ایجاد شود.میزان پیشنهادات شخصی [۱۱] حاصل از تکنیک های خوشه بندی معمولا از دیگر روشها کمتر است و در برخی از موارد میزان دقت خوشه ها از الگوریتمهای نزدیک ترین همسایه پایین تر است[6].

هرتینگ یک تکنیک مبتنی بر گراف است که در آن گره ها نشاندهنده کاربران است و یالهای بین گره ها بیانگر میزان شباهت بین دو کاربراست[1].در یک مطالعه با استفاده از داده های غیر واقعی ۱۲ ،هورتینگ پیشنهادات بهتری را نسبت به الگوریتم نزدیک ترین همسایه ارایه کرد[1].

اسکافرو همکارانش[۱۳] [26] طبقه بندی و مثالهایی را از سیستمهای پیشنهاددهنده که در تجارت الکترونیک به کار می روند و همچنین اینکه چطور آنها فضای شخصی سازی یک به یک [۱۴]را فراهم می کنند ودر همان زمان وفاداری مشتری [۱۵] را بدست می آورند،ارایه کردند.اگر چه این سیستمها در گذشته موفق بوده اند ولی استفاده گسترده از آنها کاستیهایی را نشان داده است.این کاستیها مشکلاتی مانند پراکندگی در مجموعه داده هاست[۱۶] و یا مشکلاتی که از ابعاد بالای داده ها ناشی می شود.مشکل پراکندگی در مقالات رفرنس با شماره های ۱۱و۲۳ مورد بحث محققان قرار گرفته و مشکلات ناشی از ابعاد بالای داده ها و تکنیک های کاهش بعد در مقالات رفرنس با شماره های ۴و۲۴مورد بحث قرارگرفته است.

۱-۲- آنچه در این مقاله ارایه می شود:

۱- تشریح الگوریتم پیش بینی مبتنی بر کالا و تعیین راههای متفاوت برای اجرای زیر وظیفه[۱۷] آنها

۲- ارایه یک مدل پیش محاسبه تشابه کالا برای افزایش مقیاس پذیری برخط پیشنهادات مبتنی بر کالا.

در بخش بعدی مقاله، پیشینه کوتاهی از الگوریتمهای پالایش مشارکتی ارایه می شود.در ابتدا فرایند پالایش مشارکتی تشریح شده و سپس دونگرش مبتنی بر حافظه و مبتنی بر مدل مورد بحث قرار می گیرد.در ادامه برخی از چالشهای پیش روی نگرش مبتنی بر حافظه ارایه می شود.در بخش سوم مقاله،نگرش مبتنی بر کالا ارایه شده و زیروظیفه های متفاوت الگوریتم به صورت کامل تشریح می شود.در بخش چهارم نیز نتیجه گیری مقاله و برخی از تجربیات استفاده از این الگوریتم ارایه می شود.

۲- سیستم های پیشنهاد دهنده مبتنی بر الگوریتم پالایش مشارکتی

سیستم های پیشنهاد دهنده، تکنیکهای تحلیل داده را جهت کمک به کاربران برای پیدا کردن کالاهایی که در سایتهای تجاری دوست دارند بخرند، به کار می گیرند.این سیستمها از طریق ارایه نمره احتمالی پیش بینی شده[۱۸] یا فهرستی از Top-N کالای پیشنهادی به کاربران فعال،آنان را در خرید کالاهای مورد علاقه‌شان کمک می کنند.

ارائه پیشنهاد در رابطه با خرید کالا از طرق مختلف می تواند ایجاد شود. پیشنهادات می توانند بر اساس ویژگیهای جمعیتی کاربران، کالاهای با فروش عالی یا عادات گذشته کاربران در خرید کالاها باشند. پالایش مشارکتی در حال حاضر موفق ترین تکنیک برای پیشنهاد است[19,27].

ایده اصلی الگوریتم های مبتنی برپالایش مشارکتی، فراهم نمودن پیش‌بینی یا پیشنهادات کالایی بر اساس نظرات دیگر کاربران همفکر است. نظرات کاربران می تواند به طور صریح و آشکارا از کاربر بدست آید یا به طور ضمنی از طریق بعضی ابزارها استخراج گردد.

۲-۱- نگاه کلی بر فرایند پالایش مشارکتی

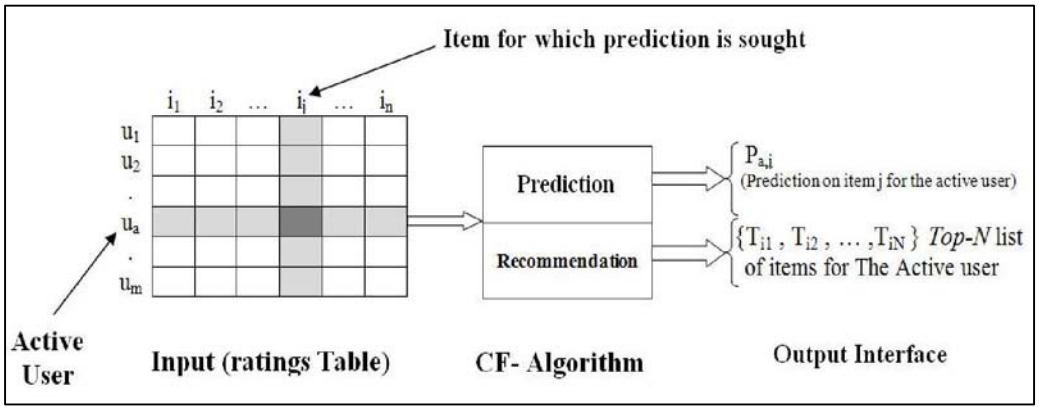
هدف الگوریتم پالایش مشارکتی، پیشنهاد کالاهای جدید یا پیش بینی مطلوبیت یک کالای معین برای یک مشتری خاص بر اساس تمایلات قبلی مشتری و نظرات دیگر مشتریان هم عقیده است. به طورکلی درسناریوی الگوریتم پالایش مشارکتی ، لیستی از m کاربر به شکل $u=\{u_1,u_2,\ldots,u_m\}$ و لیستی از n کالا مثل $I=\{i_1,i_2,\ldots,i_n\}$ داریم. هر مشتری u_i لیستی از کالاهای I_{ui} دارد که مشتری نظراتش را در مورد آنها بیان کرده است. نظرات می توانند به طور آشکار و از طریق نمره دهی مشتری به کالاها که معمولاً با مقیاس شمارشی است، اخذ شوند ویا به طور ضمنی و بر گرفته از اطلاعات مربوط به خریدهای مشتری و یا بر طبق تحلیل ثبتهای زمانبندی شده وقایع[۱۹] ویا کاوش ابرپیوندهای وبی [۲۰] و ... بدست آیند[16,28]. باید توجه داشت که I_{ui} زیر مجموعه I است و I_{ui} می تواند تهی هم باشد. در اینجا یک کاربر متمایز u_a که عضو u است، کاربر فعال نامیده می شود. وظیفه الگوریتم پالایش مشارکتی به طور مشخص، پیدا کردن کالا، متناسب باسلیقه کاربر فعال به دو شکل زیراست :

پیش بینی[۲۱]:

پیش بینی یک مقدار عددی است مثل $P_{a,j}$که شباهت کالای i_j که عضو I_{ua} نیست را برای مشتری فعال u_a بیان می کند. این مقدار پیش بینی شده با همان مقیاسی (به عنوان مثال از ۱ تا ۵) که نظرات ازکاربرفعال(u_a)اخذ می شود،بیان می شود.

پیشنهاد[۲۲]:

پیشنهاد لیستی از N کالا است؛به عبارت بهتر I_r زیر مجموعه I است، که مشتری فعال آنها را بیشتر دوست خواهد داشت. قابل توجه آنکه که لیست پیشنهادی باید بر اساس کالاهایی باشدکه قبلاً توسط مشتری فعال خریداری نشده باشد یعنی $I_r \cap I_{ua} = \Phi$. این شکل از الگوریتم های پالایش مشارکتی به عنوان پیشنهاد Top-N معروف است.



شکل 1 :فرایند پالایش مشارکتی[30]

شکل ۱ شمایی از فرایند پالایش مشارکتی را نمایش می دهد. الگوریتمهای پالایش مشارکتی، داده کالا- مشتری ورودی را به عنوان ماتریس $m \times n$ رتبه ها، A، نمایش می دهند. هر عنصر $a_{i,j}$ در A، نمره اولویت(رتبه بندی) کاربر i ام روی کالا jام را نشان می دهد. رتبه بندی افراد در یک مقیاس عددی مشخص می شود ومقدار ۰، نشان دهنده این است که کاربر هنوز رتبه بندی روی کالا نداشته است.

الگوریتمهای پالایش مشارکتی متعددی تاکنون معرفی شده اند که می توان آنها را به دو دسته اصلی مبتنی بر حافظه و مبتنی بر مدل تقسیم کرد[6]. در این قسمت تحلیلی از این الگوریتمها ارائه می شود.

الگوریتمهای پالایش مشارکتی مبتنی بر حافظه:

الگوریتمهای مبتنی بر حافظه از تمامی پایگاه داده کالا- مشتری برای ایجاد و ارایه یک پیش بینی استفاده می کند.این قبیل سیستمها از تکنیکهای آماری برای پیدا کردن مجموعه ای از کاربران که همسایه ها نامیده می شوند،استفاده می کند.این کاربران در تاریخچه رفتاری خود رفتارهای مشابهی را با کاربر هدف[۲۳] داشته اند؛ به طور مثال این همسایه ها نظرات مشابه با مشتری هدف را در خصوص کالاهای مختلف ارایه کرده ویا تمایل خودرا برای خرید مجموعه ای از کالای مشابه با خرید مشتری هدف نشان داده اند.

زمانی که یک همسایگی از مشتریان شکل می گیرد، این سیستمها از الگوریتمهای متفاوتی برای ترکیب سلیقه های همسایگان استفاده می کند تا پیش بینی یا پیشنهاد Top-N را به مشتری فعال ارایه کنند. این تکنیک ها همچنین به عنوان نزدیک ترین همسایه یا پالایش مشارکتی مبتنی بر کاربر معروف هستند و به طور گسترده در عمل مورد استفاده قرار می گیرند.

الگوریتمهای پالایش مشارکتی مبتنی بر مدل:

الگوریتمهای پالایش مشارکتی مبتنی بر مدل برای پیشنهاد کالا ابتدا یک مدلی را از نظرات مشتریان[۲۴] می سازند.الگوریتمهای حاضر در این طبقه بندی از یک روش احتمالی استفاده می کنند و با فرض وجود رتبه بندیهای مشتری موردنظردرباره سایر کالاها،از فرایند پالایش مشارکتی برای پیش بینی ارزش مورد انتظار مشتری استفاده می شود. فرایند ساخت مدل بوسیله الگوریتمهای متفاوت یادگیری ماشین مانند شبکه بیزین،خوشه بندی و روشهای مبتنی بر قاعده[۲۵] ارایه می شود.مدل شبکه بیزین یک مدل احتمالی را برای مساله پالایش مشارکتی فرمول بندی می کند[6].مدل خوشه بندی با مساله پالایش مشارکتی به مثابه مشکل کلاس بندی[۲۶] [2,6,29] رفتار می کند. روش مبتنی بر قاعده برای پیدا کردن وابستگی کالاهایی که با هم خریداری شده اندو در نهایت تولید و ارایه پیشنهاد کالا بر اساس این وابستگیها از الگوریتمهای کشف قانون وابستگی استفاده می کنند[25].

۲-۲- چالش های الگوریتمهای پالایش مشارکتی مبتنی بر کاربر

سیستم های پالایش مبتنی بر کاربر، در گذشته بسیار موفق بودند، اما استفاده وسیع از آنها برخی از چالش های بالقوه را در مورد آنها آشکار کرده است که عبارتند از :

۲-۲-۱- پراکندگی[۲۷]:

در عمل بسیاری از سیستم های پیشنهاد دهنده تجاری، برای ارزیابی مجموعه بزرگی از کالاها مورد استفاده قرار می گیرند. (به عنوان مثال Amazon.com پیشنهاد کتاب می دهد.)در این سیستم ها، حتی فعال ترین مشتریان، ممکن است کمتراز ۱٪ کالاها را خرید کرده باشند. (۱٪ از ۲ میلیون کتاب معادل ۲۰۰۰۰ کتاب می شود.) بنابراین سیستم پیشنهاددهنده بر اساس الگوریتم های نزدیکترین همسایه ممکن است قادر به ایجاد هیچ انتخابی برای کاربر خاص نباشد و در نتیجه دقت پیشنهادات پایین خواهد بود.

۲-۲-۲- مقیاس پذیری[۲۸]:

الگوریتمهای نزدیکترین همسایه ، نیازمند محاسباتی هستند که با افزایش تعداد مشتریان و هم تعداد کالاها،این محاسبات افزایش می یابند.حال با داشتن میلیونها مشتری و کالا، سیستم های معمول پیشنهاد دهنده مبتنی بر وب، از مشکل جدی مقیاس پذیری رنج می برند.

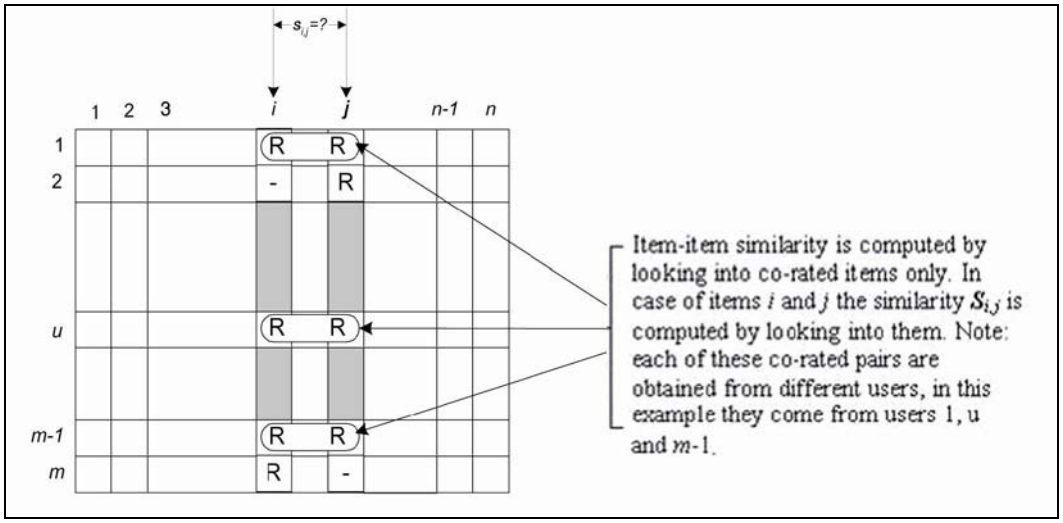
ضعف الگوریتم نزدیکترین همسایه برای پایگاه داده های پراکنده و بزرگ، محققان را بر آن داشت تا به سمت کشف الگوریتمهای جایگزین برای سیستمهای پیشنهاددهنده ، حرکت کنند.

روش اول برای گریز از مشکل پراکندگی، استفاده ازعاملهای پالایش نیمه هوشمند در سیستم است[11,23]. این عاملها با استفاده از ویژگی های ترکیب [۲۹]، هر کالا را نرخ بندی و ارزیابی می کنند. از طریق ایجاد یک مجموعه رتبه بندی شده متراکم، این عاملها به مشکل پوشش و بهبود کیفیت کمک می کنند. روش دیگر به کارگیری یک راهکار الگوریتمیک و استفاده از شاخص گذاری معنایی پنهان(LSI) برای بدست آوردن شباهت بین مشتریان و کالاها در فضای کاهش ابعاد است[24,25].

اما دراینجا به تکنیک دیگری که همانا روش مبتنی بر مدل است ومخصوصاً در بحث مقیاس پذیری بسیارموثراست،نگاهی می اندازیم. ایده مهم در این تکنیک، تحلیل ماتریس کالا- مشتری برای تعیین روابط بین کالاهای مختلف و سپس استفاده از این ارتباطها برای محاسبه رتبه پیش بینی یک جفت کالا- مشتری مفروض است. ایده این روش این است که یک مشتری علاقمند به خرید کالاهایی بوده که مشابه کالاهایی است که قبلا دوست داشته و از کالاهایی دوری می کند که مشابه کالاهایی بوده که کاربر قبلا دوست نداشته و از آنها اجتناب می ورزیده است. این تکنیک ها نیازمند تعیین همسایگی مشتریان مشابه در زمان درخواست پیشنهاد نیستند و در نتیجه آنها متمایل به تولید و ارایه پیشنهادات با سرعت بالاترهستند. تعدادی متنوعی از روشهای مختلف اعم از احتمالی[6] تاروشهای مبتنی بر تعیین همبستگی بین کالا – کالا برای محاسبه روابط بین کالاها پیشنهاد شده اند[13,15].

۳- الگوریتم پالایش مشارکتی مبتنی بر کالا

در این قسمت یک نوع از الگوریتمهای پیشنهاد مبتنی برکالا را برای تولید پیش بینی برای کاربر بررسی می کنیم. برخلاف الگوریتم پالایش مشارکتی مبتنی بر کاربرکه در قسمت دوم این مقاله بحث شد، روش مبتنی بر کالا به مجموعه ای از کالاهایی که کاربر فعال رتبه بندی کرده توجه دارد و محاسبه می کند چه مشابهت هایی با کالای هدف i دارد و سپس k کالای بیشتر مشابه $\{i_1,i_2,\ldots,i_k\}$را انتخاب می کند. در همان زمان مشابهت های $\{s_{i1},s_{i2},\ldots,s_{ik}\}$ محاسبه می شوند. زمانی که مشابه ترین کالاها پیدا می شوند، با بدست آوردن میانگین وزنی از رتبه بندی های کاربرهدف روی این کالاهای مشابه، پیش بینی به دست می آید.در این قسمت دو مرحله از این الگوریتم شامل "محاسبه مشابهت" و " تولید پیش بینی "با جزئیات بیان می شود.



شکل 2 :جداسازی کالاهای جفت رتبه بندی شده و محاسبه مشابهت[30]

۳-۱- محاسبه تشابه کالا[۳۰]:

یک گام مهم در الگوریتم های پالایش مشارکتی مبتنی برکالا محاسبه تشابه بین کالاها و سپس انتخاب شبیه ترین کالاها است. ایده اساسی در محاسبه مشابهت بین دو کالا i و jدر ابتدا، مجزا کردن مشتریانی است که هر دو این کالاها را رتبه بندی کرده اند و سپس به کاربردن یک تکنیک محاسبه مشابهتی برای تعیین مشابهت $s_{i,j}$ است. شکل ۲ این فرایند را توضیح داده است. در این شکل سطرهای ماتریس، مشتریان را نمایش می دهند وستون ها هم کالاها را به نمایش می گذارند .

راههای مختلفی برای محاسبه مشابهت بین کالاها وجود دارد. در اینجا سه روش بیان می شود :

- مشابهت براساس کسینوس
- مشابهت براساس همبستگی
- مشابهت کسینوسی تنظیم شده.

۳-۱-۱- مشابهت بر اساس کسینوس[۳۱]:

در این مورد، دو کالا به عنوان دو بردار در فضای کاربری m بعدی در نظر گرفته می شوند. مشابهت بین آنها بوسیله محاسبه کسینوس زاویه بین دو بردار، اندازه گیری می شود. در ماتریس رتبه بندی $m \times n$ ، شکل۲، مشابهت بین کالاهای i وjبوسیله $sim(i,j)$ مشخص شده و به صورت زیر است:

$$sim(i,j) = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\|_2 * \|\vec{j}\|_2} \qquad (1)$$

که علامت "." ضرب نقطه ای بین دو بردار را نمایش می دهد.

۳-۱-۲- مشابهت بر اساس همبستگی:

در این حالت، مشابهت بین دو کالای i وjاز طریق محاسبه ضریب همبستگی پیرسون[۳۲] به صورت $corr_{i,j}$ ، اندازه گیری می شود. برای ایجاد دقت محاسباتی همبستگی، ابتدا باید موارد با هم رتبه بندی شده را جدا کرد. (مثلا مواردی که مشتری هر دو کالا i وjرا رتبه بندی کرده است). مجموعه ای از مشتریان را که هر دو کالا i وjرا رتبه بندی کرده اند U درنظرمی گیریم و مشابهت همبستگی به صورت زیر بدست می آید :

$$sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_i)^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R}_j)^2}}. \qquad (2)$$

در اینجا $R_{u,i}$ رتبه بندی کاربر u روی کالا i و میانۀ $\bar{R}_i$ رتبه بندی کالای i ام است.

۳-۱-۳- مشابهت کسینوسی تنظیم شده[۳۳]:

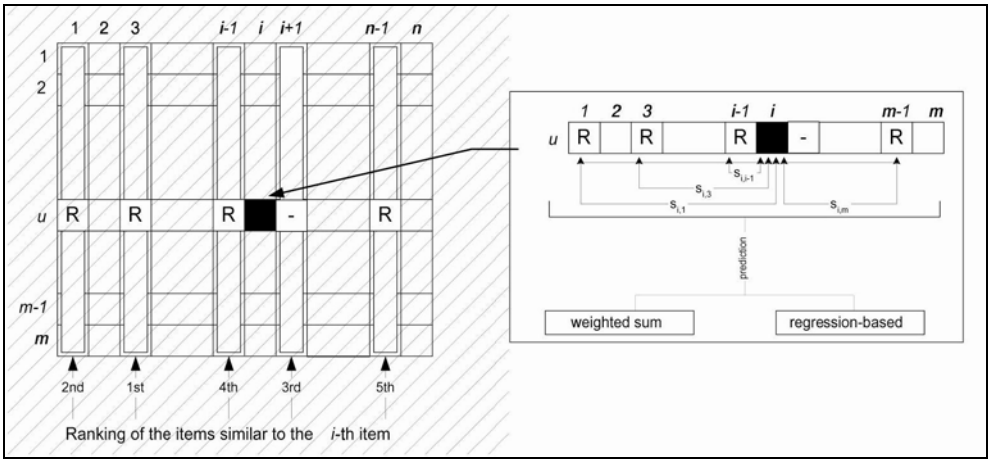
یک تفاوت اساسی بین محاسبه مشابهت در پالایش مشارکتی مبتنی برکاربر با پالایش مشارکتی مبتنی برکالا در این است که مشابهت در پالایش مشارکتی مبتنی برکاربر، در طول سطرهای ماتریس محاسبه انجام می شود اما در مورد پالایش مشارکتی مبتنی برکالا در طول ستون.

محاسبه تشابه با استفاده از روش کسینوسی در موارد مبتنی برکالا یک اشکال مهم دارد و آنهم این است که تفاوت در مقیاس رتبه بندی شده بین کاربران مختلف لحاظ نمی شود. مشابهت کسینوسی تنظیم شده این مانع را بوسیله کم کردن مقدارمیانگین کاربر متناظر با هرجفت کالای رتبه بندی شده، متعادل می کند.

فرمول مشابهت بین کالای i وjمورد استفاده به شکل زیر است :

$$sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_u)^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R}_u)^2}}. \qquad (3)$$

در اینجا میا _ رتبه بندی کاربرuام است.



شکل3 :فرایند تولید پیش بینی برای ۵ همسایه در پالایش مشارکتی مبتنی بر کالا[30]

۳-۲- محاسبه پیش بینی[۳۴]:

مهمترین مرحله در یک سیستم پالایش مشارکتی، تولید رابط خروجی[۳۵]مربوط به پیش بینی است. وقتی که مجموعه ای ازشبیه ترین کالاها براساس ابزارهای تعیین مشابهت مجزا می شوند، مرحله بعدی مشاهده رتبه بندی های مشتریان هدف و استفاده از یک تکنیک برای به دست آوردن پیش بینی ها است. در اینجا دو تکنیک بررسی می شود :

۳-۲-۱- مجموع وزنی[۳۶]:

همانطور که از نام تکنیک مشخص است این روش پیش بینی روی یک کالای i را برای کاربرu از طریق محاسبه مجموع رتبه های داده شده بوسیله مشتری روی کالاهای مشابه i ، محاسبه می کند. هر رتبه بندی بوسیله مشابهت Si,j متناظر بین کالاهای i وj وزن دهی می شود. به صورت تفصیلی این نکات در شکل۳ مشخص شده اند و پیش بینی $p_{u,i}$ به صورت زیر تعین می شود:

$$P_{u,i} = \frac{\sum_{all\ similar\ items,N}(s_{i,N} * R_{u,N})}{\sum_{all\ similar\ items,N}(|s_{i,N}|)} \qquad (4)$$

اساساً این روش سعی دارد تاکیفیت ارزیابی کاربر فعال روی کالاهای مشابه را بدست آورد. مجموع وزنی بوسیله مجموع شرایط مشابهت برای اطمینان از اینکه پیش بینی مطابق محدوده از قبل تعریف شده باشد، اندازه گیری می شود.

۳-۲-۲- رگرسیون:

این روش مشابه متد میانگین وزنی است اما به جای استفاده مستقیم رتبه بندی های کالاهای مشابه از تقریبی از رتبه بندی ها براساس مدل رگرسیون استفاده می کند. در عمل، مشابهت های محاسبه شده با استفاده از کسینوس یا روش همبستگی ممکن است گمراه کننده باشد؛بدین صورت که در حالتی که دو بردار رتبه[۳۷]، فاصله دار هستند (در مفهوم اقلیدسی) هنوز مشابهت های خیلی زیادی دارند. در این مورد استفاده از رتبه بندی های سطری به اصطلاح" کالای مشابه"، ممکن است باعث پیش بینی ضعیفی گردد. ایده اساسی، استفاده از فرمول مشابه فرمول مجموع وزنی است، اما به جای استفاده مقادیر رتبه بندی N سطری کالا مشابه این مدل ازمقدار تقریبی $R_{u,N}$ خودش، براساس مدل رگ $R'_{u,N}$ خطی استفاده می کند. اگر بردارهای کالای هدف i و کالای مشابه N بوسیله Ri و R_N مشخص شوند، مدل رگرسیون خطی می تواند به صورت زیر بیان شود :

$$\bar{R'_N} = \alpha \bar{R}_i + \beta + \epsilon \qquad (5)$$

α و β پارامترهای مدل رگرسیون هستند که با حرکت بر روی هر دو بردارهای رتبه بندی تعیین می شوند. € خطای پارامتر مدل رگرسیون است.

۴- نتیجه گیری:

سیستمهای پیشنهاددهنده فناوری قدرتمند و جدیدی هستند که از طریق پایگاه داده مشتریان ، باعث ایجاد ارزش افزوده درکسب و کارها می گردد.این سیستم ها مشتریان را در پیدا کردن کالاهایی که می خواهند خریداری نمایند،کمک می نمایند. سیستمهای پیشنهاددهنده با توانمند ساختن مشتریان در خرید کالاهای مورد علاقه شان به مشتریان سود می رسانند و در مقابل، این سیستمها کسب و کارها را با ایجاد فروش بیشتر یاری می رسانند. سیستمهای پیشنهاددهنده به سرعت در حال تبدیل شدن به یکی از ابزارهای حیاتی سایتهای تجارت الکترونیک در محیط وب هستند.برای بهبود مساله مقیاس پذیری در سیستمهای پیشنهاددهنده نیاز به تکنولوژیهای جدید است.

در این مقاله، الگوریتمهای پالایش مشارکتی که در سیستمهای پیشنهاددهنده به کار می رود تشریح شد.نتایج تحقیقات [30]و تجربیات شرکتهایی که این از این الگوریتم استفاده کرده اند[31]،نشان از موفقیت این الگوریتمها دارد.

مراجع:

[1] Aggarwal, C. C., Wolf, J. L., Wu K., and Yu, P. S. (1999). *Horting Hatches an Egg: A New Graph-theoretic Approach to Collaborative Filtering*. In Proceedings of the ACM KDD'99 Conference. San Diego, CA. pp. 201- 212.

[2] Basu, C., Hirsh, H., and Cohen, W. (1998). *Recommendation as Classification: Using social and Content-based Information in Recommendation*. In Recommender System Workshop '98. pp. 11-15.

[3] Berry, M. W., Dumais, S. T., and O'Brian, G. W. (1995).*Using Linear Algebra for Intelligent Information Retrieval*. SIAM Review, 37(4), pp. 573-595.

[4] Billsus, D., and Pazzani, M. J. (1998).*Learning Collaborative Information Filters*. In Proceedings of ICML '98. pp. 46-53.

[5] Brachman, R., J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., and Simoudis, E. 1996.*Mining Business Databases*, Communications of the ACM, 39(11), pp. 42-48, November.

[6] Breese, J. S., Heckerman, D., and Kadie, C. (1998). *Empirical Analysis of Predictive Algorithms for Collaborative Filtering*. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, pp. 43-52.

[7] Cureton, E. E., and D'Agostino, R. B. (1983). *Factor Analysis: an Applied Approach*, Lawrence Erlbaum associates pubs, Hillsdale, NJ.

[8] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). *Indexing by Latent Semantic Analysis*, Journal of the American Society for Information Science, 41(6), pp. 391-407.

[9] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., Eds. (1996). *Advances in Knowledge Discovery and Data Mining*, *AAAI press/MIT press*.

[10] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. (1992). *Using Collaborative Filtering to Weave an Information Tapestry*, Communications of the ACM, December.

[11] Good, N., Schafer, B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., and Riedl, J. (1999). *Combining Collaborative Filtering With Personal Agents for Better Recommendations*, In Proceedings of the AAAI-'99 conference, pp 439-446.

[12] Herlocker, J., Konstan, J., Borchers, A., and Riedl, J. (1999). *An Algorithmic Framework for Performing Collaborative Filtering*, In Proceedings of ACM SIGIR'99. ACM press.

[13] Herlocker, J. (2000). *Understanding and Improving Automated Collaborative Filtering Systems*, Ph.D. Thesis, Computer Science Dept., University of Minnesota.

[14] Hill, W., Stead, L., Rosenstein, M., and Furnas, G. (1995). R*ecommending and Evaluating Choices in a Virtual Community of Use*. In Proceedings of CHI '95.

[15] Karypis, G. (2000). *Evaluation of Item-Based Top-N Recommendation Algorithms*, Technical Report CS-TR-00- 46, Computer Science Dept., University of Minnesota.

[16] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. (1997). *GroupLens: Applying Collaborative Filtering to Usenet News*, Communications of the ACM, 40(3), pp. 77-87.

[17] Ling, C. X., and Li C. (1998). *Data Mining for Direct Marketing: Problems and Solutions*, In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, pp. 73-79.

[18] Peppers, D., and Rogers, M. (1997). *The One to One Future: Building Relationships One Customer at a Time*, Bantam Doubleday Dell Publishing.
[19] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). *GroupLens: An Open Architecture for Collaborative Filtering of Netnews*. In Proceedings of CSCW '94, Chapel Hill, NC.
[20] Resnick, P., and Varian, H. R. (1997). *Recommender Systems,* Communications of the ACM, 40(3).
[21] Reichheld, F. R., and Sasser Jr., W. (1990). *Zero Defections: Quality Comes to Services,* Harvard Business School Review, 1990(5): pp. 105-111.
[22] Reichheld, F. R. (1993). *Loyalty-Based Management*, Harvard Business School Review, 1993(2): pp. 64-73.
[23] Sarwar, B., M., Konstan, J. A., Borchers, A., Herlocker, J., Miller, B., and Riedl, J. (1998). *Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System*, In Proceedings of CSCW '98, Seattle, WA.
[24] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. (2000). *Application of Dimensionality Reduction in Recommender System–A Case Study*, In ACM WebKDD 2000 Workshop.
[25] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. (2000). *Analysis of Recommendation Algorithms for E-Commerce*, In Proceedings of the ACM EC'00 Conference. Minneapolis, MN. pp. 158-167
[26] Schafer, J. B., Konstan, J., and Riedl, J. (1999). *Recommender Systems in E-Commerce*. In Proceedings of ACM E-Commerce 1999 conference.
[27] Shardanand, U., and Maes, P. (1995). *Social Information Filtering: Algorithms for Automating 'Word of Mouth'*, In Proceedings of CHI '95. Denver, CO.
[28] Terveen, L., Hill, W., Amento, B., McDonald, D., and Creter, J. (1997). *PHOAKS: A System for Sharing Recommendations*, Communications of the ACM, 40(3). pp. 59-62.
[29] Ungar, L. H., and Foster, D. P. (1998) *Clustering Methods for Collaborative Filtering*, In Workshop on Recommender Systems at the 15th National Conference on Artificial Intelligence.
[30] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2001), *Item-based Collaborative Filtering Recommendation Algorithms*, Proceedings of the 10th international conference on World Wide Web Hong Kong, pp: 285 – 295.
[31] Linden, G., Smith, B., and York, J. (Jan/Feb 2003), *Amazon.com recommendations: item-to-item collaborative filtering*, In IEEE Internet computing, Volume: 7, pp: 76- 80.

[1] - Item-based Algorithms
[2] - User-based Algorithms
[3] - Online
[4] - Usenet news
[5] - Bayesian networks
[6] - Clustering
[7] - Horting graph
[8] - Training set
[9] - Off-line
[10] - Nearest neighbor
[11] - Personal recommendation
[12] - Synthetic data
[13] - Schafer et al.
[14] - One-to-one personalization
[15] - Customer loyalty
[16] - Sparsity in data set
[17] - Subtasks
[18] - predicted likeliness score
[19] - Analyzing timing logs
[20] -Web mining hyperlinks
[21] - Prediction
[22] - Recommendation
[23] - Target user
[24] - User rating
[25] - Rule – based methods
[26] - Cclassification
[27] - Sparsity
[28] - Scalability
[29] - Syntactic Features
[30] - Item Similarity Computation
[31] - Cosine-based similarity
[32] - Pearson Correlation
[33] - Adjusted Cosine Similarity
[34] - Prediction Computation
[35] - Output Interface

[36] - Weighted sum
[37] - Rating Vector