چهارمین
همایش ملی تجارت الکترونیکی
The 4th e-Commerce
CONFERENCE
تهران- ۳ و ۴ آذرماه ۱۳۸۶

کاربرد سیستمهای پیشنهاد دهنده در تجارت الکترونیک

نفیسه شبیب، دکتر محمد علی نعمت بخش، سیما عقیلی دهکردی

چکیده

رشد بی سابقه تکنولوژی جدید اینترنت در سالهای اخیر، باعث ایجاد برنامه های کاربردی بسیار زیادی در زمینه تجارت الکترونیکی شده است. وجود برنامه های کاربردی در زمینه B2B و B2C نیاز به ارتباط موثر بین ماشینها را دارد. یکی از مهمترین برنامه های کاربردی سیستمهای پیشنهاد دهنده می باشد. سیستم‌های پیشنهاد دهنده یک نوع ویژه از سیستم‌های فیلتر اطلاعات است، که در آن آیتم‌ها را، بر اساس اینکه چه آیتمی برای کاربر جذاب است، از یک مجموعه بزرگ از آیتم‌ها و کاربران فیلتر می‌کنند. در این مقاله سیستمهای پیشنهاد دهنده را بررسی نموده و نسل حاضر از متدهای سیستمهای پیشنهاد دهنده را که به سه دسته کلی تقسیم می شوند را مطرح می نماییم.

کلید واژه ها: سیستمهای پیشنهاد دهنده[1]، فیلتر همبستگی[2]، فیلتر مبتنی بر محتوا[3]، فیلتر ترکیبی[4]

[1] Recommender System
[2] Collaborative Filtering
[3] Content-Base Filtering
[4] Hybrid Filtering

1- مقدمه

سیستمهای پیشنهاد دهنده با اولین ظهورشان در زمینه فیلتر همبستگی حوزه تحقیقاتی مهمی در اواسط دهه 1990 را فراهم نمودند[1]. دردهه های اخیر دو بخش صنعت و دانشگاه دستاوردهای جدیدی در زمینه سیستم های پیشنهاد دهنده توسعه داده اند؛ با این وجود علاقه مندی به این بخش هنوز در سطح بالایی است. زیرا حوزه تحقیقاتی غنی بوده و نیاز به برنامه های کاربردی فراوانی به منظور کمک به کاربران که با حجم زیادی از اطلاعات مواجه هستند به منظور شخصی سازی اطلاعات پیشنهادی وجود دارد. مثالهایی از چنین برنامه های مثل سیستم پیشنهاد دهی کتاب، سی دی و دیگر محصولات سایت آمازون [2] یا پیشنهاد فیلم توسط شرکت Movie lens [3] و پیشنهاد اخبار توسط سایت adaptiveinfo.com [4] می باشند. در این مقاله راههای متنوعی از توسعه توانایی های سیستمهای پیشنهاد دهنده را بیان می کنیم.

2- مروری بر سیستم های پیشنهاد دهنده

دراواسط دهه 1990، زمانیکه محققان تحقیقاتشان را در زمینه سیستمهای پیشنهاد دهنده آغاز نمودند این تحقیقات بطور روشنی روی ساختارهای نرخ گذاری متمرکز بود. دراغلب فرمول ها، مسایل پیشنهاد دهی با تخمین نرخ که معمولا توسط کاربران داده می شد، کاهش می یافت. تخمین ها معمولا براساس نرخهای داده شده به آیتمها توسط کاربر واطلاعات دیگرکه بصورت قراردادی وجود داشت زده می شد. دراینصورت آیتمهایی به کاربر پیشنهاد می شد که بالاترین نرخ را داشتند. بطور قراردادی مسئله پیشنهاد دهی بصورت زیر فرمول بندی می کنیم.

دراین روش C را به عنوان همه کاربران و S را به عنوان همه آیتمهای ممکن مثل کتاب، فیلم، رستوران و ...که می تواند به کاربر پیشنهاد شود درنظر می گیریم. فضای S از آیتمهای ممکن می تواند خیلی بزرگتر باشد و دامنه ای بین صد، هزاران یا حتی میلیونها آیتم در هر کاربردی باشد؛ مثل پیشنهاد کتاب یا سی دی که می تواند در بعضی موارد فضایی میلیونی داشته باشد. تابع سودمندی u، سودمندی آیتم s را برای کاربر c بیان می کند. مجموعه کل سفارشات را با R نشان داده که بصورت $C\times S\rightarrow R$ تعریف می نماییم. سپس برای هر کاربر $c \in C$ آیتمی مثل $s'\rightarrow S$ راکه سودمندی کاربررا ماکزیموم نماید بصورت قراردادی بصورت زیر تعریف می نماییم.

$$\forall c \in C, \quad s'_c = \arg\max_{s \in S} u(c, s). \tag{1}$$

درسیستمهای پیشنهاد دهنده، سودمندی یک آیتم معمولا بوسیله نرخ گذاری مشخص می شود و بیان می کند چگونه کاربر خاصی آیتمی را دوست دارد. به عنوان مثال john Doe به فیلم Harry Potter امتیاز 7 از 10 را می دهد. بهرحال درحالت کلی، سودمندی می تواند یک تابع اختیاری شامل تابع سود باشد. سودمندی u می تواند توسط کاربر، به عنوان آنچه اغلب به عنوان نرخ تعریف شده توسط کاربر[1] یا بوسیله برنامه، تحت تابع سودمندی مبتنی برسود[2]، تعریف گردد. هر جزئی از فضای کاربر C می تواند با پروفایلی که شامل مشخصه های متنوعی همچون سال، جنسیت، درآمد، وضعیت ازدواج و ... است، مشخص شود. مورد ساده ای از پروفایل می تواند جزء ساده ای مثل شماره کاربر[3] باشد. درهرصورت هر جزیی از فضای آیتم Sبا مجموعه ای از مشخصه ها تعریف شده است. برای مثال دربرنامه پیشنهاد دهی فیلم، جاییکه S مجموعه ای از فیلم هاست، هر فیلم فقط می تواند با ID مشخص شود، در صورتیکه می تواند توسط مشخصه های دیگری مثل نام کارگردان، سال تولید، بازیگران فیلم و ... مشخص شود.

[1] User-defined rating
[2] Profit-benefit
[3] User ID

برای مثال در برنامه پیشنهاد دهی فیلم مثل آنچه در سایت Movielens.org می باشد، کاربران بصورت اولیه به زیر مجموعه ای از فیلمهایی که قبلا دیده اند، نرخی را اختصاص می دهند. یک مثال از ماتریس نرخ دهی کاربر-آیتم برای پیشنهاد فیلم در جدول 1 مشخص شده است؛ جاییکه نرخها بین 1 تا 5 هستند و نماد Ø برای بعضی از نرخها درجدول1 به معنای اینست که کاربران نرخی به فیلم نداده اند، بنابراین سیستم پیشنهاد دهنده بایستی بتواند فیلم های نرخ داده نشده و انواع مناسب پیشنهاد بر اساس این نرخها تخمین زده و پیش بینی نماید

	K-PAX	Life of Brian	Memento	Notorious
Alice	4	3	2	4
Bob	Ø	4	5	5
Cindy	2	2	4	Ø
David	3	Ø	5	2

جدول1 ماترس نرخ دهی برای سیستم پیشنهاد دهنده فیلم

سیستمهای پیشنهاد دهنده براساس تخمین نرخ و چگونگی پیشنهادات به سه دسته تقسیم می شود.

1) **فیلترکننده براساس محتوا**: دراین روش پیشنهادات براساس آیتمهایی که کاربر در گذشته خریده است، ارائه می شود.

2) **فیلتر همبستگی:** دراین روش براساس شباهت رفتاری والگوهای عملکردی کاربرانی که شباهت های رفتاری و الگوهای مشابهی با کاربر فعلی در گذشته داشته اند، پیشنهادات ارائه می شود.

3) فیلتر ترکیبی: این روش ترکیبی از دو روش بالاست و با استفاده از ترکیب متدها، تاحدودی مشکلات فیلتر هبستگی و فیلتر کننده بر اساس محتوا را بر طرف می کند.

3- فیلتر کننده بر اساس محتوا

در متد مبتنی بر محتوا سودمندی$u(c,s)$ از آیتم s برای کاربرc، از روی سودمندی $u(c,s_i)$ بوسیله کاربرc به آیتمهای S $\in s_i$که s_i ها شبیه آیتم S هستند ، تخمین زده می شود. برای مثال در برنامه پیشنهاد دهنده فیلم به کاربر c، این سیستم سعی می کند اشتراکات بین فیلم ها مثل نام کارگردان، نوع فیلم، موضوع فیلم، بازیگران خاص و... که کاربرc به آنها نرخ بالایی داده تشخیص دهد و در اینصورت فیلم های که درجه تشابه بالاتری، با اولویت های مشتری دارند را پیشنهاد نماید.

3-1 بررسی فرایند فیلتر مبتنی بر محتوا

روش مبتنی برمحتوا، برای پیشنهاد دهی ریشه در بازیابی اطلاعات [5] و فیلترنمودن اطلاعات [6] دارد. ازآنجاییکه پیشرفت های قابل توجه و چشمگیری توسط بازیابی اطلاعات و انجمن های فیلترینگ در زمینه سیستم های مبتنی بر متن انجام شده است، بسیاری از سیستمهای پیشنهاد دهنده، روی آیتمهایی مبتنی بر اطلاعات متنی مثل اسناد، آدرس وب سایتها و متن پیامهای خبری یوزنت متمرکز شده اند. روشهای بازیابی اطلاعات به شیوه سنتی از پروفایل کاربر به منظور اطلاع از اولویت های مشتری و نیازهای او استفاده می کنند. اطلاعات پروفایل یا بصورت واضحی از طریق پرسشنامه یا بصورت ضمنی از رفتار تراکنشات بدست می آید. بصورت فرمال$content(s)$ ازیک پروفایل آیتم که مجموعه ای از مشخصه های آیتم s می باشد بوسیله استخراج مجموعه ای از خصوصیات آیتم s، برای مشخص نمودن تناسب آیتم برای اهداف پیشنهاد، استخراج می شود. همانطور که قبلا توضیح داده شد، سیستمهای مبتنی بر محتوا اغلب برای آیتمهای مبتنی بر متن طراحی شده اند. محتوا دراین سیستمها معمولا با کلمات کلیدی[1] مشخص می شود. اهمیت وسودمندی کلمهk_i درسندd_j باوزن w_{ij} تعریف می شود.

[1] Key words

یکی از بهترین روش‌های اندازه‌گیری وزن کلمات کلیدی در بازیابی اطلاعات استفاده از اندازه‌گیری تعداد تکرارواژه/ معکوس تعداد تکرار سند[1] است که به صورت زیر تعریف می‌شود: فرض کنیم که N تعداد کل متن‌هایی باشد که می‌توانند به کاربر پیشنهاد داده شوند و همچنین کلمه کلیدی K_j در n_i تا از آنها وجود دارد. به علاوه فرض کنیم که $f_{i,j}$ تعداد دفعات تکرار کلمه کلیدی K_i در متن d_j است .سپس $TF_{i,j}$ ، تکرار واژه کلیدی K_i در متن d_j است، که به صورت زیر تعریف می‌شود. که ماکزیمم تعداد تکرارهای $f_{z,j}$ برای تمام کلمات کلیدی K_z که در متن d_j ظاهر شده است، محاسبه می‌شود.

$$TF_{i,j} = \frac{f_{i,j}}{\max_z f_{z,j}} \quad (2)$$

هنگامی که یک کلمه کلیدی در متن‌های زیادی تکرار می‌شود دیگر آن کلمه کلیدی برای نمایش تفاوت بین متن‌ها نمی-تواند زیاد مفید باشد و با کمک آن کلمه کلیدی نمی‌توان تشخیص داد کدام متن مناسب و کدام متن غیرمناسب است. به همین دلیل اغلب از اندازه‌گیری معکوس تعداد تکرار داکیومنت[2] (IDF_i) در ترکیب با تعداد تکرار واژه‌های ساده ($TF_{i,j}$) استفاده می‌کنند. معکوس تعداد تکرار داکیومنت برای کلمه کلیدی K_i معمولاً به صورت زیر تعریف می‌شود:

$$IDF_i = \log \frac{N}{n_i} \quad (3)$$

سپس، وزن TF-IDF برای کلمه کلیدی k_i در متن d_j به صورت زیر تعریف می‌شود:

$$w_{i,j} = TF_{i,j} \times IDF_i \quad (4)$$

و محتوای متن d_j نیز مانند زیر است:

$$Content(d_j) = (w_{1j}, \ldots w_{kj}).$$

همان گونه که قبلاً توضیح دادیم، سیستم‌های مبتنی بر محتوا آیتم‌های مشابه با آنچه که کاربر قبلاً انتخاب کرده است، به کاربر پیشنهاد می دهد[7]. آیتم‌های کاندیدای گوناگون، با آیتمی که قبلاً توسط کاربر نرخ‌گذاری شده است، مقایسه می‌شوند و بهترین آیتم که با آیتم موردنظر هماهنگ است، پیشنهاد داده می‌شود. و برای فرمال‌تر کردن این منظور، یک پروفایل مربوط به کاربر با عنوان پروفایل‌محتواگرا (c)[3] در نظر می‌گیریم. این پروفایلی شامل علایق و ترجیحات کاربر c است. این پروفایل با استفاده از تحلیل محتواهایی از آیتم‌هایی که قبلاً کاربر مشاهده کرده و یا نرخ‌گذاری کرده است، به دست می‌آید. معمولاً با کمک تکنیک‌های تحلیل کلمات کلیدی از بازیابی اطلاعات بدست می‌آید. به عنوان نمونه، پروفایل c شامل یک بردار از وزن-های ($w_{c1}, w_{c2} \ldots, w_{ck}$) است، که هر وزن w_{ci} نشان دهنده اهمیت کلمه کلیدی k_i برای کاربر c است. در سیستم مبتنی بر محتوا ، تابع سودمندی معمولاً به صورت زیر تعریف می‌شود.

$$u(c,s) = score(ContentBasedProfile(c), Content(s)). \quad (5)$$

برای محاسبه دقیق‌تر تابع سودمندی، هم باید از پروفایل مربوط به آیتم‌ها و هم پروفایل مربوط به کاربر استفاده کرد و بردارهای w_c و w_s را برای کلمات کلیدی از فرمول TF-IDF محاسبه کرد. و برای تعیین میزان سودمندی آیتم s برای کاربر

[1] Term frequency/inverse document frequency (TF-IDF)
[2] Inverse Document Frequency(IDF_i)
[3] Content-Based Profile(c)

c باید میزان تشابه پروفایل آیتم s و پروفایل کاربر c را اندازه بگیریم که برای محاسبه این تشابه می‌توانیم از فرمول اندازه گیری تشابه کسینوسی[1] استفاده کرد:

$$u(c,s) = \cos(\vec{w}_c, \vec{w}_s) = \frac{\vec{w}_c \cdot \vec{w}_s}{||\vec{w}_c||_2 \times ||\vec{w}_s||_2} = \frac{\sum_{i=1}^{K} w_{i,c} w_{i,s}}{\sqrt{\sum_{i=1}^{K} w_{i,c}^2} \sqrt{\sum_{i=1}^{K} w_{i,s}^2}}, \quad (6)$$

که k در اینجا نشان‌دهنده تعداد کلمات کلیدی در کل سیستم است.

برای مثال اگر کاربرc مقالات آنلاین زیادی درمورد زیست شناسی می خواند، تکنیکهای مبتنی بر متن می تواند مقالات دیگری در زمینه زیست شناسی ومقالات مرتبط دیگری در زمینه ژنتیک و ... را به وی پیشنهاد نماید. بنابراین پروفایل مبتنی بر متنc، تعریف شده توسط بردارw_c واژه هایی مثل k_i با وزن w_{ic} را نمایش می دهد. نتیجتا سیستم های پیشنهاد دهنده ای که از کسینوس یا ابزارهای مشابه استفاده می کنند، تابع سودمندی$u(c,s)$ به مقالات sی که واژه زیست شناسی درآنها وزن بالاتری (w_s بالاتر) دارد، سودمندی بیشتر و به مقالاتی که واژه زیست شناسی وزن کمتری دارد، سودمندی کمتری اختصاص می دهند.

3-2- محدودیت‌های الگوریتم مبتنی بر محتوا

بهرحال سیستمهای پیشنهاددهنده مبتنی بر متن محدودیتهایی رادارند [3] که در ادامه مطرح می کنیم.

3-2-1- محدودیت آنالیز محتوا

الگوریتم مبتنی بر محتوا توسط مشخصه های اشیا محدود می شود و بنابراین نیازبه مجموعه کافی و متناسبی از مشخصه هاست تا سیستم پیشنهاد دهنده یا بصورت دستی یا اتوماتیک مشخصه ها را تجزیه نموده وخصوصیات آنها را به منظور پیشنهاد آیتم کشف نماید. تکنیک‌های بازیابی اطلاعات تا زمانی که آیتم‌ها به صورت متنی باشند به خوبی می‌تواند مشخصه ها را استخراج کنند، اما انواع دیگری از آیتم‌ها به طورذاتی با مسأله استخراج اتوماتیک مشخصه ها مشکل دارند. به عنوان مثال، متدهای استخراج اتوماتیک مشخصه ها در مورد داده‌های مالتی مدیا مانند عکس‌های گرافیکی، داده‌های صوتی و داده-های ویدیویی، بامشکلات فراوانی مواجه هستند و علاوه براین اغلب امکان‌پذیر نیست که تا خصوصیات را به صورت دستی وارد کنیم [8].

3-2-1-1- زیاد اختصاصی کردن

در این سیستم‌ها، سیستم فقط آیتم‌هایی شبیه به پروفایل کاربر را به کاربر پیشنهاد می‌دهد، در نتیجه کاربر فقط آیتم‌هایی مشابه به آنچه را که قبلاً نرخ‌گذاری کرده است می‌تواند مشاهده کند

3-2-1-2- مسئله اولین کاربر

سیستم مبتنی بر محتوا فقط زمانی می‌تواند به کاربر پیشنهاد قابل اعتماد بدهد که کاربر به تعداد کافی آیتم از قبل نرخ داده باشد و سیستم بتواند علایق کاربر را تشخیص دهد. بنابراین، یک کاربر جدید، که نرخ‌های کمتری به آیتمها داده است، سیستم پیشنهاد دهنده قادر نیست پیشنهادات درستی ارائه کند.

[1] Cosine similarity

4- فیلتر همبستگی

برخلاف متد سیستمهای پیشنهاد دهنده مبتنی بر محتوا، سیستم پیشنهاد دهنده همبستگی یا فیلتر همبستگی سعی می کند سودمندی آیتمها را برای کاربر خاصی براساس آیتمهای نرخ داده شده قبلی توسط دیگر کاربران پیش بینی نماید. بطوررسمی سودمندی u(c,s) از آیتم s برای کاربر c بوسیله سودمندی $u(c_j,s)$ متعلق به آیتم s بوسیله دیگر کاربران $c_j \in C$که مشابه کاربر c هستند، تخمین زده می شود. برای مثال دریک برنامه پیشنهاددهی فیلم، به منظور پیشنهاد فیلم به کاربر، سیستم پیشنهاد دهی همبستگی سعی می کند کاربران مشابه که سلیقه مشابهی در انتخاب فیلم به کاربرc داشته اند را پیدا کرده و فقط فیلم هایی که کاربران مشابه دوست دارند به کاربر c پیشنهاد دهد. سیستم های فیلتر همبستگی در صنعت ودانشگاه توسعه داده شده اند. سیستم Grundy[8] اولین سیستم پیشنهاد دهی است که از کلیشه[1] (رفتاریکنواخت) به عنوان مکانیزمی برای ساخت مدلهایی از کاربران براساس مقدار محدودی از اطلاعات فردی کاربراستفاده نمود. با استفاده از کلیشه ها، سیستم Grundy مدلهای کاربرفردی را ساخته واز آنها در پیشنهاد کتب مربوط به هر کاربر استفاده نمود. بعد از آن سیستم Tapestry برای پیشنهاد دهی به کاربران همفکر و متجانس استفاده شد. سیستمهایGroup lens[9] و Ringo اولین سیستمهایی بودند که از الگوریتم های فیلتر همبستگی بصورت اتوماتیک استفاده نمودند.

4-1- بررسی فرایند فیلتر همبستگی

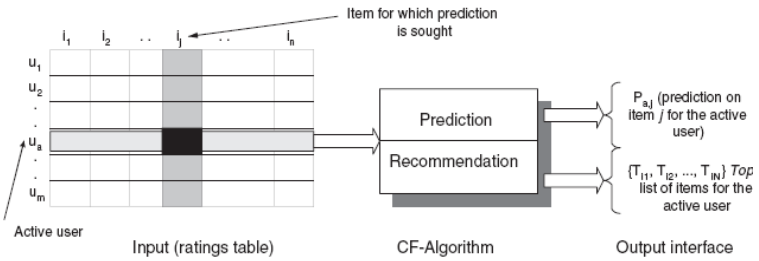
هدف فیلتر همبستگی یافتن کاربرانی است که هم عقیده با کاربر جدید هستند،سپس آیتم های مورد علاقه آنها را به کاربر جدید پیشنهاد می کند.در این فرایند یک لیست از mکاربر و یک لیست هم از n آیتم ساخته می شود :

$$\mathcal{U} = \{u_1, u_2, \ldots, u_m\} \ , \ \mathcal{I} = \{i_1, i_2, \ldots, i_n\}$$

هر کاربر u_i لیستی از آیتم هایی را دارد که نظر خود را در مورد آنها اعلام کرده است : I_{u_i} .نظرات u_i می تواند صریحاً بصورت فاکتورهای نرخ که از کاربر گرفته می شود ، بیان شود و یا بصورت ضمنی از نحوه تعاملات و خریدهای قبلی کاربر استخراج شود .به هرحال داریم : $I_{u_i} \subseteq \mathcal{I}$.

فرض کنید کاربر جدیدی بنام $u_a \in \mathcal{U}$ وارد شده است .

- در ابتدا در مرحله پیشگویی[2] میزان علاقه کاربر به آیتم هایی که تا کنون خرید نکرده است، بر اساس ماتریس آیتم-کاربر[3] تخمین زده می شود: $P_{a,j}$
- سپس در مرحله پیشنهاد لیستی از N تا آیتم ای که کاربر بیشترین علاقه را به آن دارد ، ساخته می‌شود.شکل 2-1روش کار را نشان می دهد.



شکل2-1: فرایند فیلتر همبستگی

[1] Stereotype
[2] Prediction
[3] user-item

در یک دسته‌بندی که در [10] توضیح داده شده است الگوریتم‌های CF به دو نوع اصلی 1)مبتنی بر حافظه[1] یا مبتنی بر هیوریستیک 2)مبتنی بر مدل[2] تقسیم بندی می شوند.

4-2- الگوریتم مبتنی بر حافظه

الگوریتم مبتنی بر حافظه[10] ذاتا هیوریستیک است و پیش بینی بر اساس مجموعه کل آیتمهای نرخ داده شده بوسیله کاربرانجام می شود. ارزش نرخ نامشخص $r_{c,s}$ برای کاربر c وآیتم s معمولا با تجمع نرخ دیگر کاربران مشابه s تعیین می شود.

$$r_{c,s} = \operatorname*{aggr}_{c' \in \hat{C}} r_{c',s}, \quad (9)$$

جاییکه $\hat{C}$ مشخص می کند مجموعه ای از N کاربراست که شباهت زیادی با کاربرc که به آیتم s نرخ داده اند دارد. (N می تواند دامنه ای بین 1 تا تعداد کل کاربران داشته باشد). در بعضی از مثالها تجمع[3] بصورت زیر می باشد.

$$\begin{aligned} &(a)\ r_{c,s} = \frac{1}{N}\sum_{c' \in \hat{C}} r_{c',s}, \\ &(b)\ r_{c,s} = k\sum_{c' \in \hat{C}} sim(c,c') \times r_{c',s}, \\ &(c)\ r_{c,s} = \bar{r}_c + k\sum_{c' \in \hat{C}} sim(c,c') \times (r_{c',s} - \bar{r}_{c'}) \end{aligned} \quad (10)$$

مضروب فیه k بعنوان فاکتورنرمالسازی بکار می رود توسط فرمول $k = 1/\sum_{c' \in \hat{C}} |sim(c,c')|$ محاسبه می شود و میانگین نرخ کاربرc تحت فرمول زیر محاسبه می گردد.

$$\bar{r}_c = (1/|S_c|)\sum_{s \in S_c} r_{c,s}, \text{ where } S_c = \{s \in S | r_{c,s} \neq \oslash\} \quad (11)$$

دریک مورد ساده، تجمع می تواند میانگین ساده ای همانند آنچه در فرمول(10a)در نظر گرفته ایم، محاسبه گردد. بهرحال بیشتر دستاوردهای تجمع ازمجموع وزنهای نشان داده شده در فرمول (10b)استفاده می کنند. ابزارهمسانی بین بین کاربر c و c' که با $sim(c,c')$ نشان می دهیم، اندازه گیری فاصله بکاررفته، بعنوان وزن است که هر چه کاربرcو c' به یکدیگر شبیه تر باشند نرخ وزنی بیشتر $r_{c',s}$ در پیش بینی $r_{c,s}$ بکار می رود. نکته اینجاست که $sim(c,c')$ یک هیوریستیک مصنوعی است که به منظورتوانمند نمودن تفاوت بین سطوح کاربران مشابه (جفت های مشابه[4] و همسایگان نزدیک[5] به هر کاربر) معرفی شده و در زمان مشابه پروسیجرهای تخمین نرخ را ساده می کند. همانطورکه در (10b) نشان داده شده است، برنامه های پیشنهاد دهی مختلف می توانند از ابزارهای همسانی خودشان استفاده نمایند تازمانیکه محاسبات با فاکتورنرمالسازی k نرمال شود. دوابزاراندازه گیری همسانی در زیر توضیح داده خواهد شد. یک مسئله در استفاده از جمع وزنها در فرمول (10b) اینست که ممکن است کاربران مختلف از مقیاسهای نرخ دهی مختلفی استفاده نمایند. درفرمول (10c) جمع وزنها متعادل شده و محدودیت ها را در نظر گرفته است. دراین روش به جای قدرمطلق ارزش وزنها ازنرخ انحراف از میانگین وزنها استفاده شده است. راه دیگری که بر نرخ دهی در مقیاسهای مختلف غلبه کرده است، استفاده از فیلترینگ مبتنی بر اولویتها[6][11] است که روی پیش بینی اولویتهای نسبی کاربر به جای ارزش گذاری مطلق تمرکز نموده است. دستاوردهای متنوعی برای محاسبه تشابه $sim(c,c')$ بین کاربران درسیستمهای پیشنهاددهنده فیلترهمبستگی بدست امده است. دراغلب این روشها شباهت بین دو کاربر روی نرخ گذاری آنها از آیتمهای که کاربران به ان نرخ داده اند مبتنی است. دوروش عمومی عبارتند از:

[1] Memory-base
[2] Model-base
[3] Aggregation
[4] Closest peers
[5] Nearest neighbors
[6] Preference base filtering

- محاسبه تشابه مبتنی بر کسینوس[1]
- محاسبه تشابه مبتنی بر همبستگی[2]

برای نشان دادن آنها، S_{xy} را مجموعه ای ازآیتمهای وابسته توسط دو کاربر x,y درنظر می گیریم $S_{xy} = \{s \in S | r_{x,s} \neq \oslash \ \& \ r_{y,s} \neq \oslash\}$.درسیستم فیلتر محتوا S_{xy} اساسا به عنوان نتیجه میانی محاسبات نزدیکترین همسایه کاربر x درنظر می گیریم. اغلب ازروشهای مستقیم توسط تقاطع مجموعه S_yوS_x محاسبه می کنیم. بهرحال بعضی از متدها مثل تئوری گراف[3] در فیلتر همبستگی می تواند نزدیکترین همسایه به xرا بدون محاسبه همه S_{xy} برای کاربرy بدست آورد.

درروش تشابه مبتنی بر همبستگی از ضریب همبستگی پیرسون[4] برای محاسبه همسانی استفاده می شود[11].

$$sim(x,y) = \frac{\sum_{s \in S_{xy}} (r_{x,s} - \bar{r}_x)(r_{y,s} - \bar{r}_y)}{\sqrt{\sum_{s \in S_{xy}} (r_{x,s} - \bar{r}_x)^2 \sum_{s \in S_{xy}} (r_{y,s} - \bar{r}_y)^2}} \quad (12)$$

در روش تشابه مبتنی برکسینوس[10] دو کاربر x وy به عنوان دو بردار در فضای m بعدی جاییکه $m=|S_{xy}|$ می باشد، در نظر گرفته می شود. سپس شباهت دو بردار می تواند بوسیله محاسبه زاویه کسینوس بین آنها محاسبه شود.

$$sim(x,y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}||_2 \times ||\vec{y}||_2} = \frac{\sum_{s \in S_{xy}} r_{x,s} r_{y,s}}{\sqrt{\sum_{s \in S_{xy}} r_{x,s}^2} \sqrt{\sum_{s \in S_{xy}} r_{y,s}^2}} \quad (13)$$

جاییکه $\vec{x} \cdot \vec{y}$ مشخص می کند نقطه حاصلضرب بین دو بردار $\vec{y}$و$\vec{x}$ می باشد.

روش دیگر برای اندازه گیری همسانی بین کاربران ابزار تفاوت میانگین مربعات[5]است که در[11] توضیح داده شده است. تفاوت بین روشها مختلف سیستمهای پیشنهاد دهنده روش محاسبه همسانی کاربران به منظور کارانمودن بیشتر می باشد. استراتژی عمومی محاسبه همسانی کاربر$sim(x,y)$ (شامل محاسبهS_{xy}) و محاسبه مجدد آنها تنها زمانیکه شبکه ای از کاربران مشابه درزمان کوتاهی تغییر نکند؛ و هر گاه کاربر برای پیشنهاد درخواست می کند، نرخها بصورت موثر با استفاده از همسانی های محاسبه شده قبلی به کاربر پیشنهاد می شود.

در هر دو روش مبتنی بر محتوا و فیلتر همبستگی برای بازیابی اطلاعات از ابزار تشابه مبتنی بر کسینوس، استفاده می شود. بهرحال در سیستم پیشنهاد دهی براساس محتوا، ابزار تشابه بین بردارهای وزنی TF-IDF است، درحالیکه درسیستم فیلتر همبستگی تشابه بین بردارهای نرخهای کاربران مشخص می شود. برای توسعه تکنیک های فیلتر همبستگی، روشهای اصلاح در بهبود کارایی مثل فرکانس معکوس کاربر[6]، پیش بینی وزنهای اکثریت[7]، default voting پیشنهاد شده است. برای مثال مثال default voting [10] روش توسعه مبتنی بر حافظه است. دراین روش ابزارتشابه روی تقاطع مجموعه آیتمها می باشد؛

[1] Cosine base similarity
[2] Correlation base similarity
[3] Graph-theoretic
[4] Pearson correlation coefficient
[5] Mean squared difference
[6] Inverse user frequency
[7] Weighted-majority prediction

بعنوان مثال روی مجموعه نرخهای داده شده بوسیله کاربر x وy تعدادی نرخ پیش فرض برای نرخهای مفقود قرار می دهیم و بدین ترتیب دقت پیش بینی بهبود می یابد

4-3- الگوریتم مبتنی بر مدل

درمقابل الگوریتمهای مبتنی بر حافظه، الگوریتمهای مبتنی بر مدل قرار دارند. همانطور که در[10]بیان شده این الگوریتمها برای پیش بینی نرخها از یک مدل یادگیری استفاده می کنند. برای مثال [10] روشی براساس احتمال برای الگوریتم های فیلترهمبستگی پیشنهاد کرد. دراین روش نرخهای ناشناخته بصورت زیر محاسبه می شوند.

$$r_{c,s} = E(r_{c,s}) = \sum_{i=0}^{n} i \times \Pr(r_{c,s} = i | r_{c,s'}, s' \in S_c) \quad (14)$$

درآن ارزش نرخها را اعداد صحیحی بین 0 تاn فرض می کنیم و بیان می کنیم که احتمالاً کاربرc نرخ خاصی را به آیتم s مطابق آنچه قبلانرخ داده اند، نرخ می دهد. برای محاسبه این احتمال[10] دومدل احتمال که عبارتندازمدلهای دسته بندی یا کلاستر[1]و شبکه های بیزین[2] را پیشنهاد کرد. مدل دسته بندی یا کلاستر کاربران متجانس را درکلاسهای دسته بندی می کند. برای ارائه عضویت کاربردر کلاس، فرض می شود که نرخهای کاربر مستقل باشند و این ساختار مدل ساده بیزین[3] می باشد. تعداد کلاسها و پارامترهای مدل از داده آموخته می شوند.

درمدل شبکه بیزین هرآیتم در دامنه را به عنوان گره ای از ساختار بیزین، جاییکه نرخ هر گره به آیتم های دیگر شبیه است، درنظر می گیرد. هردو ساختارشبکه و احتمالات شرطی از داده آموخته می شوند. یکی از محدودیتهای این روش این است که هرکاربرعضو یک گروه است درحالیکه ممکن است بعضی از برنامه های پیشنهاد دهی قابلیت عضویت یک کاربر در چندین گروه را داشته باشند. برای مثال در برنامه پیشنهاد کتاب یک کاربر ممکن است به خاطرکارش به موضوعی مثل برنامه نویسی علاقه مند باشد، درحالیکه برای اوقات فراغتش کتابی در مورد آشپزی بخواهد.تفاوت اصلی بین تکنیکهای فیلترهمبستگی مبتنی بر مدل و روشهای مبتنی برحافظه یا هیوریستیک اینست که تکنیکهای مبتنی بر مدل نرخ سودمندی را به شیوه مبتنی بر هیوریستیک محاسبه نمی کنند، بلکه آنها براساس یادگیری از داده ها در آمار و تکنیکهای یادگیری ماشین می باشند.

4-4- محدودیت‌های الگوریتم فیلتر همبستگی

بهر حال الگوریتم فیلتر همبستگی محدودیتهای خاص خودش را دارد؛ که در ادامه به آنها می پردازیم.

4-4-1-مقیاس‌پذیری

درالگوریتم های فیلترهمبستگی مبتنی بر کاربر زمانی که کاربران صدها یا هزاران تن باشد به خوبی پاسخ می‌دهد اما امروزه تجارت الکترونیکی به سرعت در حال گسترش است و تعداد کاربران به بیشتر از میلیون‌ها تن رسیده است و این سیستم‌ها دیگر پاسخ‌گو نیستند زیرا در این سیستم‌ها محاسبات به صورت برخط[4] محاسبه می شود و اگر حجم اطلاعات زیاد باشد زمان

[1] Cluster models
[2] Bayesian networks
[3] Naive Bayesian model
[4] online

می کنند[1] بسیار طولانی می‌شود که دیگر قابل قبول نیست. برای رفع این مشکل از الگوریتم بر اساس آیتم استفاده

4-4-2-پراکندگی ماتریس نرخ

منظور از ماتریس نرخ، ماتریسی است با ابعاد $n \times m$ ، که n تعداد کاربران و m تعداد آیتم‌هاست و هر عنصر $i \times j$ در این ماتریس نشان‌دهنده نرخی است که توسط کاربر i برای آیتم j تخمین زده شده است. و به طور معمول کاربران با کمتر از 1% از آیتم‌های موجود در یک وب‌سایت سروکار دارند و فقط به این آیتم‌ها نرخ می‌دهند ونتیجه آن داشتن یک ماتریس بزرگ است که بیشتر عناصر آن تهی است. و این منجر می‌شود که جستجو در این ماتریس مشکل شود. و نتیجتا میزان درستی و صحت در این سیستم‌ها پایین می آید. این مسئله در الگوریتم فیلتر ترکیبی[2] تا حدودی بر طرف می شود.

4-4-3-اولین نرخ و اولین کاربر

مشکل دیگر هنگامی رخ می‌دهد که آیتمی تازه به سیستم وارد شده باشد در این صورت آن آیتم توسط کاربری نرخ داده نشده است، بنابراین آیتم نمی‌تواند انتخاب شود و به کاربران پیشنهاد داده شود. اگر از دید دیگر به این مسئله نگاه کنیم کاربری تازه به سیستم وارد شده باشد و به هیچ کالایی نرخی نداده باشد در این صورت، سیستم نمی‌تواند علایق کاربر را تشخیص دهد و آیتمی به کاربر پیشنهاد دهد.

5- الگوریتم ترکیبی

برخی از سیستمهای پیشنهاد دهنده از روش دیگری که ترکیبی از دو روش مبتنی بر محتوا و فیلتر همبستگی است، استفاده می نمایند تا محدودیتهای دو روش قبلی را کاهش دهند. راههای مختلفی برای ترکیب دو روش پیشنهاد شده است که درادامه دسته بندی می کنیم.

متدهای ترکیبی را بر اساس دسته بندی دیگری به شکل زیر دسته بندی می کنند.

- وزندار[3]: نتایج(نرخ یا امتیاز) چندین تکنیک پیشنهاد دهنده باهم ترکیب می‌شوند تا یک پیشنهاد ساده تولید شود.
- راهگزینی[4]: در این روش با توجه شرایط جاری سیستم یکی از روش‌های پیشنهاد دهنده را انتخاب می‌کند.
- آمیخته[5]: پیشنهاد از چندین سیستم پیشنهاددهنده متفاوت که در یک زمان نمایش داده شده اند، ایجاد می‌شود.
- ترکیب خصوصیات[6]: خصوصیات از منابع داده پیشنهاد دهنده های متفاوت باهم در یک الگوریتم ساده قرار می-گیرند.
- آبشار[7]: سیستم پیشنهادات دیگر سیستم‌ها را پالایش می‌کند.
- افزایش خصوصیات[8]: خروجی یک تکنیک به عنوان خصوصیت ورودی سیستم دیگر استفاده می‌شود.
- فرا سطح[9]: مدلی که یک سیستم یادگرفته به عنوان ورودی دیگران استفاده می‌شود.

---

[1] response time
[2] Hybrid filtering
[3] Weighted
[4] Switching
[5] mixed
[6] Feature combination
[7] Cascade
[8] Feature augmentation
[9] Meta-level

۵- نتیجه گیری

کاربرد اصلی سیستم‌های پیشنهاد دهنده در وب سایت‌های بزرگ برای پیشنهاد کالا به کاربران است. همانطور که قبلا توضیح داده شد، سیستمهای پیشنهاد دهنده بر اساس روش به سه دسته مبتنی بر محتوا، فیلتر همبستگی و ترکیبی تقسیم بندی می شوند و بر اساس نوع تکنیک پیشنهاد دهنده به دو دسته مبتنی بر هیوریستیک یا مبتنی بر مدل تقسیم شده اند. نتیجتا می توان گفت مهم ترین کاربرد سیستمهای پیشنهاد دهنده درکاربردهای تجارت الکترونیک می باشد. در این مقاله، همانند اغلب مقالات در زمینه فناوری اطلاعات محدودیت ها و مشکلاتی وجود داشت. یکی از مشکلات تحقیق، فقدان یک سیستم پایه تجارت الکترونیکی در ایران برای بررسی عملکرد الگوریتمهای سیستمهای پیشنهاد بود.

منابع

[1] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, “Recommending and Evaluating Choices in a Virtual Community of Use,” roc. Conf. Human Factors in Computing Systems, 1995.
[2] G. Linden, B. Smith, and J. York, “Amazon.com Recommendations Item-to-Item Collaborative Filtering,” IEEE Internet Computing, Jan./Feb. 2003.
[3] B.N. Miller, I. Albert, S.K. Lam, J.A. Konstan, and J. Riedl, “MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System,” Proc. Int’l Conf. Intelligent User Interfaces, 2003.
[4] N. Belkin and B. Croft, “Information Filtering and Information Retrieval,” Comm. ACM, vol. 35, no. 12, pp. 29-37, 1992
[5] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison-Wesley, 1999.
[6] C. Basu, H. Hirsh, and W. Cohen, “Recommendation as Classification: Using Social and Content-Based Information in Recommendation,” Recommender Systems. Papers from 1998 Workshop, Technical Report WS-98-08, AAAI Press 1998.
[7] K. Lang, “Newsweeder: Learning to Filter Netnews,” Proc. 12th Int’l Conf. Machine Learning, 1995. [57] W.S. Lee, Collaborative Learning for Recommender Systems,”Proc. Int’l Conf. Machine Learning, 2001.
[8] U. Shardanand and P. Maes, “Social Information Filtering: Algorithms for Automating ‘Word of Mouth’,” Proc. Conf. Human Factors in Computing Systems, 1995.
[9] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, and J. Riedl, “GroupLens: Applying Collaborative Filtering to Usenet News,” Comm. ACM, vol. 40, no. 3, pp. 77-87, 1997.
[10] J.S. Breese, D. Heckerman, and C. Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering,” Proc. 14th Conf. Uncertainty in Artificial Intelligence, July 1998.
[11] W.W. Cohen, R.E. Schapire, and Y. Singer, “Learning to Order Things,” J. Artificial Intelligence Research, vol. 10, pp. 243-270, 1999.