

## کاربرد سیستمهای پیشنهاد دهنده در تجارت الکترونیک

نفیسسه 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 می کنند. 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<sup>۱</sup>، فیلتر همبستگی<sup>۲</sup>، فیلتر مبتنی بر محتوا<sup>۳</sup>، فیلتر ترکیبی<sup>۴</sup>

---

<sup>۱</sup> Recommender System

<sup>۲</sup> Collaborative Filtering

<sup>۳</sup> Content-Base Filtering

<sup>۴</sup> 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]. در دهه 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](http://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

در اواسط دهه 1990، زمانیکه محققان تحقیقاتشان را در زمینه سیستمهای پیشنهاد دهنده آغاز نمودند این تحقیقات بطور روشنی روی ساختارهای نرخ گذاری متمرکز بود. در اغلب 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 و اطلاعات دیگر که بصورت قراردادی وجود داشت زده می شد. در اینصورت آیتمهایی به کاربر پیشنهاد می شد که بالاترین نرخ را داشتند. بطور قراردادی مسئله پیشنهاد دهی بصورت زیر فرمول بندی می کنیم.

در این 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 در نظر 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  $s' \rightarrow S$  را که سودمندی کاربر را ماکزیموم نماید بصورت قراردادی بصورت زیر تعریف می نماییم.

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

در سیستمهای 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. بهر حال در حالت 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<sup>1</sup> یا بوسیله برنامه، تحت تابع سودمندی مبتنی بر سود<sup>2</sup>، 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<sup>3</sup> باشد. در هر صورت هر جزئی از فضای آیتم  $S$  با مجموعه ای از مشخصه ها تعریف شده است. برای مثال در برنامه پیشنهاد دهی فیلم، جایکه  $S$  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.

<sup>1</sup> User-defined rating

<sup>2</sup> Profit-benefit

<sup>3</sup> 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، نرخ 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  $\emptyset$  برای بعضی از نرخها در جدول 1 به معنای اینست که کاربران نرخى 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	$\emptyset$	4	5	5
Cindy	2	2	4	$\emptyset$
David	3	$\emptyset$	5	2

جدول 1 ماتریس نرخ دهی برای سیستم پیشنهاد دهنده فیلم

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

- 1) **فیلتر کننده بر اساس محتوا:** در این روش پیشنهادات بر اساس آیتمهایی که کاربر در گذشته خریده است، ارائه می شود.
- 2) **فیلتر همبستگی:** در این روش بر اساس 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 همبستگی 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 کاربر  $c$ ، از روی سودمندی  $u(c,s_i)$  بوسیله کاربر  $c$  به آیتمهای  $S$   $s_i \in S$  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 کاربر  $c$  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 بر محتوا، برای پیشنهاد دهی ریشه در بازیابی اطلاعات [5] و فیلتر نمودن اطلاعات [6] دارد. از آنجاییکه 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  $content(s)$  از یک 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 در این سیستمها معمولا با کلمات کلیدی<sup>1</sup> مشخص می شود. اهمیت و سودمندی کلمه  $k_i$  درسند  $d_j$  با وزن  $W_{ij}$  تعریف می شود.

<sup>1</sup> Key words

یکی از بهترین روش‌های اندازه‌گیری وزن کلمات کلیدی در بازیابی اطلاعات استفاده از اندازه‌گیری تعداد تکرار واژه/ معکوس تعداد تکرار سند<sup>۱</sup> 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 نمی‌تواند 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 داکيومنت<sup>۲</sup> ( $IDF_i$ ) در ترکیب با تعداد تکرار واژه‌های ساده ( $TF_{i,j}$ ) استفاده می‌کنند. معکوس تعداد تکرار داکيومنت 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:

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

همان گونه که قبلاً توضیح دادیم، سیستم‌های مبتنی بر محتوا آیت‌های مشابه با آنچه که کاربر قبلاً انتخاب کرده است، به کاربر پیشنهاد می‌دهد [7]. آیت‌های کاندیدای گوناگون، با آیت‌های قبلاً توسط کاربر نرخ‌گذاری شده است، مقایسه می‌شوند و بهترین آیت که با آیت مورد نظر هماهنگ است، پیشنهاد داده می‌شود. و برای فرمال‌تر کردن این منظور، یک پروفایل مربوط به کاربر با عنوان پروفایل محتواگرا ( $c$ )<sup>۳</sup> در نظر می‌گیریم. این پروفایلی شامل علائق و ترجیحات کاربر  $c$  است. این پروفایل با استفاده از تحلیل محتوایی از آیت‌هایی 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 وزن‌های ( $w_{c1}, w_{c2}, \dots, w_{ck}$ ) 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 آیت‌ها 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 آیت  $s$  برای کاربر

<sup>1</sup> Term frequency/inverse document frequency (TF-IDF)

<sup>2</sup> Inverse Document Frequency (IDF<sub>i</sub>)

<sup>3</sup> Content-Based Profile(c)

C باید میزان تشابه پروفایل آیتم S و پروفایل کاربر C را اندازه بگیریم که برای محاسبه این تشابه می توانیم از فرمول اندازه گیری تشابه کسینوسی<sup>1</sup> استفاده کرد:

$$u(c, s) = \cos(\vec{w}_c, \vec{w}_s) = \frac{\vec{w}_c \cdot \vec{w}_s}{\|\vec{w}_c\|_2 \times \|\vec{w}_s\|_2}$$

$$= \frac{\sum_{i=1}^K w_{i,c} w_{i,s}}{\sqrt{\sum_{i=1}^K w_{i,c}^2} \sqrt{\sum_{i=1}^K w_{i,s}^2}}$$

(6)

که k در اینجا نشان دهنده تعداد کلمات کلیدی در کل سیستم است. برای مثال اگر کاربر C مقالات آنلاین زیادی در مورد زیست شناسی می خواند، تکنیکهای مبتنی بر متن می تواند مقالات دیگری در زمینه زیست شناسی و مقالات مرتبط دیگری در زمینه ژنتیک و ... را به وی پیشنهاد نماید. بنابراین پروفایل مبتنی بر متن C، تعریف شده توسط بردار  $W_c$  و واژه هایی مثل  $k_i$  با وزن  $w_{i,c}$  را نمایش می دهد. نتیجتاً سیستم های پیشنهاد دهنده ای که از کسینوس یا ابزارهای مشابه استفاده می کنند، تابع سودمندی  $u(C,S)$  به مقالات S که واژه زیست شناسی در آنها وزن بالاتری ( $w_s$  بالاتر) دارد، سودمندی بیشتر و به مقالاتی که واژه زیست شناسی وزن کمتری دارد، سودمندی کمتری اختصاص می دهند.

### 3-2-2- محدودیت های الگوریتم مبتنی بر محتوا

بهرحال سیستمهای پیشنهاددهنده مبتنی بر متن محدودیتهایی را دارند [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 نیازه مجموعه کافی و متناسبی از مشخصه هاست تا سیستم پیشنهاد دهنده یا بصورت دستی یا اتوماتیک مشخصه ها را تجزیه نموده و خصوصیات آنها را به منظور پیشنهاد آیتم کشف نماید. تکنیک های بازیابی اطلاعات تا زمانی که آیتمها به صورت متنی باشند به خوبی می تواند 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 داده های مالی مدیا مانند عکس های گرافیکی، داده های صوتی و داده های ویدیویی، بامشکلات فراوانی مواجه هستند و علاوه براین اغلب امکان پذیر 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 می دهد، در نتیجه کاربر فقط آیتمهایی مشابه به آنچه را که قبلاً نرخ گذاری کرده است می تواند مشاهده کند

#### 3-2-1-2- مسئله اولین کاربر

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

<sup>1</sup> 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. بطور رسمی سودمندی  $u(c,s)$  از آیتم  $s$  برای کاربر  $c$  بوسیله سودمندی  $u(c,s)$  متعلق به آیتم  $s$  بوسیله دیگر کاربران  $c_j \in C$  که مشابه کاربر  $c$  هستند، تخمین زده می شود. برای مثال در یک 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 کاربر  $c$  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 و دانشگاه توسعه داده شده اند. سیستم Grundy [8] اولین سیستم پیشنهاد دهی است که از کلیشه<sup>1</sup> (رفتاریکخواخت) به عنوان مکانیزمی برای ساخت مدلهایی از کاربران براساس مقدار محدودی از اطلاعات فردی کاربر استفاده نمود. با استفاده از کلیشه ها، سیستم Grundy مدلهای کاربر فردی 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. سیستمهای Group 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  $m$  کاربر و یک لیست هم از  $n$  آیتم ساخته می شود:

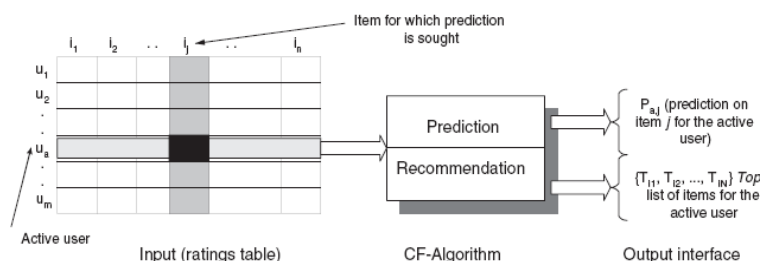
$$U = \{u_1, u_2, \dots, u_m\}, I = \{i_1, i_2, \dots, i_n\}$$

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 هر حال داریم:  $I_{u_i} \subseteq I$ .

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

- در ابتدا در مرحله پیشگویی<sup>2</sup> میزان علاقه کاربر به آیتم هایی که تا کنون خرید نکرده است، بر اساس ماتریس آیتم-کاربر<sup>3</sup> 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 می شود. شکل 2-1- آروش کار را نشان می دهد.



شکل 2-1: فرایند فیلتر همبستگی

<sup>1</sup> Stereotype

<sup>2</sup> Prediction

<sup>3</sup> user-item

در یک دسته‌بندی که در [10] توضیح داده شده است الگوریتم‌های CF به دو نوع اصلی (1) مبتنی بر حافظه<sup>1</sup> یا مبتنی بر هیوریستیک (2) مبتنی بر مدل<sup>2</sup> 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 کاربر انجام می شود. ارزش نرخ نامشخص  $r_{c,s}$  برای کاربر C و آیتm S معمولا با تجمع نرخ دیگر کاربران مشابه S تعیین می شود.

$$r_{c,s} = \text{aggr } r_{c',s}, \quad c' \in C \quad (9)$$

جاییکه C مشخص می کند مجموعه ای از N کاربر است که شباهت زیادی با کاربر C که به آیتm S نرخ داده اند دارد. (N می تواند دامنه ای بین 1 تا تعداد کل کاربران داشته باشد). در بعضی از مثالها تجمع<sup>3</sup> بصورت زیر می باشد.

$$(a) \quad r_{c,s} = \frac{1}{N} \sum_{c' \in C} r_{c',s}$$

$$(b) \quad r_{c,s} = k \sum_{c' \in C} \text{sim}(c, c') \times r_{c',s}$$

$$(c) \quad r_{c,s} = \bar{r}_c + k \sum_{c' \in C} \text{sim}(c, c') \times (r_{c',s} - \bar{r}_{c'})$$

(10)

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

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

در یک مورد ساده، تجمع می تواند میانگین ساده ای همانند آنچه در فرمول (10a) در نظر گرفته ایم، محاسبه گردد. بهرحال بیشتر دستاوردهای تجمع از مجموع وزنه‌های نشان داده شده در فرمول (10b) استفاده می کنند. ابزار همسانی بین کاربر C و C' که با  $\text{sim}(c, c')$  نشان می دهیم، اندازه گیری فاصله بکاررفته، بعنوان وزن است که هر چه کاربر C' و C به یکدیگر شبیه تر باشند نرخ وزنی بیشتر  $r_{c',s}$  در پیش بینی  $r_{c,s}$  بکار می رود. نکته اینجاست که  $\text{sim}(c, c')$  یک هیوریستیک مصنوعی است که به منظور توانمند نمودن تفاوت بین سطوح کاربران مشابه (جفت های مشابه<sup>4</sup> و همسایگان نزدیک<sup>5</sup> به هر کاربر) معرفی شده و در زمان مشابه پروسیجرهای تخمین نرخ را ساده می کند. همانطور که در (10b) نشان داده شده است، برنامه های پیشنهادی مختلف می توانند از ابزارهای همسانی خودشان استفاده نمایند تا زمانی که محاسبات با فاکتور نرمالسازی k نرمال شود. دوازده اندازه گیری همسانی در زیر توضیح داده خواهد شد. یک مسئله در استفاده از جمع وزنه‌ها در فرمول (10b) اینست که ممکن است کاربران مختلف از مقیاسهای نرخ دهی مختلفی استفاده نمایند. در فرمول (10c) جمع وزنه‌ها متعادل شده و محدودیت ها را در نظر گرفته است. در این روش به جای قدرمطلق ارزش وزنه‌ها از نرخ انحراف از میانگین وزنه‌ها استفاده شده است. راه دیگری که بر نرخ دهی در مقیاسهای مختلف غلبه کرده است، استفاده از فیلترینگ مبتنی بر اولویتهای<sup>6</sup> [11] است که روی پیش بینی اولویتهای نسبی کاربر به جای ارزش گذاری مطلق تمرکز نموده است. دستاوردهای متنوعی برای محاسبه تشابه  $\text{sim}(c, c')$  بین کاربران در سیستمهای پیشنهاددهنده فیلتر همبستگی بدست آمده است. در اغلب این روشها شباهت بین دو کاربر روی نرخ گذاری آنها از آیتمهای که کاربران به آن نرخ داده اند مبتنی است. دوروش عمومی عبارتند از:

<sup>1</sup>Memory-base

<sup>2</sup>Model-base

<sup>3</sup>Aggregation

<sup>4</sup>Closest peers

<sup>5</sup>Nearest neighbors

<sup>6</sup>Preference base filtering



- محاسبه تشابه مبتنی بر کسینوس<sup>1</sup>
- محاسبه تشابه مبتنی بر همبستگی<sup>2</sup>

برای نشان دادن آنها،  $S_{xy}$  را مجموعه ای از آیتمهای وابسته توسط دو کاربر  $x, y$  در نظر می گیریم  $S_{xy} = \{s \in S | r_{x,s} \neq 0 \ \& \ r_{y,s} \neq 0\}$ . در سیستم فیلتر محتوا  $S_{xy}$  اساساً به عنوان نتیجه میانی محاسبات نزدیکترین همسایه کاربر  $x$  در نظر می گیریم. اغلب از روشهای مستقیم توسط تقاطع مجموعه  $S_x$  و  $S_y$  محاسبه می کنیم. بهر حال بعضی از متدها مثل تئوری گراف<sup>3</sup> در فیلتر همبستگی می تواند نزدیکترین همسایه به  $x$  را بدون محاسبه همه  $S_{xy}$  برای کاربر  $y$  بدست آورد. در روش تشابه مبتنی بر همبستگی از ضریب همبستگی پیرسون<sup>4</sup> 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 بر کسینوس [10] دو کاربر  $x$  و  $y$  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  $\vec{x}$  و  $\vec{y}$  می باشد. روش دیگر برای اندازه گیری همسانی بین کاربران ابزار تفاوت میانگین مربعات<sup>5</sup> است که در [11] توضیح داده شده است. تفاوت بین روشها مختلف سیستمهای پیشنهاد دهنده روش محاسبه همسانی کاربران به منظور کارآمدن بیشتر می باشد. استراتژی عمومی محاسبه همسانی کاربر  $sim(x, y)$  (شامل محاسبه  $S_{xy}$ ) و محاسبه مجدد آنها تنها زمانیکه شبکه ای از کاربران مشابه در زمان 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. بهر حال در سیستم پیشنهاد دهی براساس محتوا، ابزار تشابه بین بردارهای وزنی TF-IDF است، درحالیکه در سیستم 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<sup>6</sup>، پیش بینی وزنها<sup>7</sup> اکثریت،  $default\ voting$  پیشنهاد شده است. برای مثال  $default\ voting$  [10] روش توسعه مبتنی بر حافظه است. در این روش ابزار تشابه روی تقاطع مجموعه آیتمها می باشد؛

<sup>1</sup> Cosine base similarity

<sup>2</sup> Correlation base similarity

<sup>3</sup> Graph-theoretic

<sup>4</sup> Pearson correlation coefficient

<sup>5</sup> Mean squared difference

<sup>6</sup> Inverse user frequency

<sup>7</sup> Weighted-majority prediction



بعنوان مثال روی مجموعه نرخهای داده شده بوسیله کاربر X و Y 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 تا n فرض می کنیم و بیان می کنیم که احتمالاً کاربر C نرخ خاصی را به آیتm S مطابق آنچه قبلانرخ داده اند، نرخ می دهد. برای محاسبه این احتمال [10] دومدل احتمال که عبارتندازمدلهای دسته بندی یا کلاستر<sup>1</sup> و شبکه های بیزین<sup>2</sup> 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<sup>3</sup> می باشد. تعداد کلاسهها و پارامترهای مدل از داده آموخته می شوند.

درمدل شبکه بیزین هرآیتm در دامنه را به عنوان گره ای از ساختار بیزین، جاییکه نرخ هر گره به آیتm های دیگر شبیه است، درنظر می گیرد. هر دو 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 بخوهد. 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- محدودیت های الگوریتم فیلتر همبستگی

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

##### 4-4-1- مقیاس پذیری

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

<sup>1</sup> Cluster models

<sup>2</sup> Bayesian networks

<sup>3</sup> Naive Bayesian model

<sup>4</sup> online

#### 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 نرخ 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 و نتیجه آن داشتن یک ماتریس بزرگ است که بیشتر عناصر آن تهی است. و این منجر می‌شود که جستجو در این ماتریس مشکل شود. و نتیجتاً میزان درستی و صحت در این سیستم‌ها پایین می‌آید. این مسئله در الگوریتم فیلتر ترکیبی<sup>2</sup> تا حدودی بر طرف می‌شود.

#### 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 نرخ نداده باشد در این صورت، سیستم نمی‌تواند علائق 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 می‌نمایند تا محدودیتهای دو روش قبلی را کاهش دهند. راههای مختلفی برای ترکیب دو روش پیشنهاد شده است که در ادامه دسته بندی می‌کنیم.

متمدهای ترکیبی را بر اساس دسته بندی دیگری به شکل زیر دسته بندی می‌کنند.

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

<sup>1</sup> response time

<sup>2</sup> Hybrid filtering

<sup>3</sup> Weighted

<sup>4</sup> Switching

<sup>5</sup> mixed

<sup>6</sup> Feature combination

<sup>7</sup> Cascade

<sup>8</sup> Feature augmentation

<sup>9</sup> Meta-level

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 در کاربردهای 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," Proc. 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.