نشـــریه بین المللـــی مهنـــدسی صنـــایع و مدیـــــریت تولیـــد

شمـــــاره ۱، جلـــد ۲۳، بهـــار ۱۳۹۱
صفحـــــه ۱۳-۲

نشریه بین المللی مهندسی صنایع و مدیریت تولید دانشگاه علم و صنعت ایران

ISSN: 2008-4870

http://IJIEPM.iust.ac.ir/

سیستم پیشنهادگر هوشمند برای خرده‌فروشی اینترنتی با استفاده از نقشهٔ خودسازمانده و قواعد انجمنی بر اساس الگوهای جمعیت‌شناختی مشتریان

شهاب مسیّبیان، عبّاس کرامتی* و وحید خطیبی

کلمات کلیدی

سیستمهای پیشنهادگر،
خوشه‌بندی، نقشهٔ خودسازمانده،
قواعد انجمنی،
خرده‌فروشی اینترنتی

چکیده:

امروزه به دلیل گستردگی رقابت در دنیای تجارت الکترونیکی، روشهای مؤثر در جذب مشتریان از اهمیّت ویژه‌ای برخوردار شده‌اند. یکی از این روشها، بکارگیری سیستمهای پیشنهادگر در وبگاه‌های تجاری است تا بدین‌ترتیب امکان استخراج علایق مشتریان و پیشنهاد مناسب‌ترین محصولات به آنها میسّر گردد. در این مقاله، مدل جدیدی برای سیستمهای پیشنهادگر ارایه شده است که به کمک آن می‌توان بخش‌بندی بازار و مشتری را به شیوهٔ کارآمدتری انجام داده و در نتیجه پیشنهادات بهتری به مشتری ارایه داد. بدین‌منظور از روشهای داده‌کاوی همچون خوشه‌بندی و قواعد انجمنی استفاده شده است، به‌طوریکه در فاز اول خوشه‌بندی مشتریان بر اساس مشخصه‌های جمعیّت‌شناختی سن، جنسیّت، شغل و تحصیلات انجام شده است که در آن تعداد خوشه‌ها با استفاده از الگوریتم نقشهٔ خودسازمانده (SOM) مشخص شده و سپس خوشه‌ها با الگوریتم K میانگین (K-Means) ایجاد گردیده‌اند. در فاز دوم با استفاده از قواعد انجمنی در هر خوشه، نقشه‌ای معتبر انتخاب شده و بر اساس آن به مشتریان آن خوشه، پیشنهادات مناسب گوناگونی ارائه شده است. برای بررسی کارایی مدل پیشنهادی، از آن در تحلیل داده‌های یک وبگاه تجاری ایرانی برای پیشنهاددهی به مشتریان استفاده گردیده است که نتایج مناسبی از خوشه‌بندی و ارایهٔ پیشنهادات حاصل شد.

۱. مقدمه

با گسترش روزافزون بکارگیری اینترنت در امور مختلف زندگی انسان، شیوهٔ انجام مبادلات تجاری از طریق این شبکه به یکی از مهمّترین مباحث مطرح در آن تبدیل شده است که مورد توجّه بسیاری از محقّقان، سازمانها و مشتریان قرار گرفته است [۱]. بر این اساس، تلاشهای فراوانی برای راه‌اندازی سیستمهای خرید اینترنتی صورت گرفته، شیوهٔ نوینی برای انجام خرید در محیط مجازی رقم زده شده است، به طوری که پیامد صرفه‌جویی قابل ملاحظه در زمان و هزینه نسبت به خرید سنتی را در پی داشته است [۲].

با تکامل بازاریابی، هم‌اکنون شرکتهای موفق مشتری‌محوری را با شعار بازاریابی فردبه‌فرد[2] مورد توجّه قرار داده‌اند. در تجارت الکترونیک نیز خدمات متناسب با سلایق کاربران، شخصی‌سازی[3] شده و امروزه به روالی رایج در این عرصه تبدیل شده است [۳].

تاریخ وصول: ۸۹/۸/۸
تاریخ تصویب: ۹۰/۳/۲

*نویسنده مسئول مقاله: دکتر عبّاس کرامتی، عضو هیات علمی گروه مهندسی صنایع، پردیس دانشکده‌های فنی، دانشگاه تهران keramati@ut.ac.ir
شهاب مسیّبیان، کارشناس ارشد مهندسی فنّاوری اطلاعات، دانشکدهٔ فنی و مهندسی، دانشگاه تربیت مدرس؛ mosayebian@modares.ac.ir .
وحید خطیبی، دانشجوی دکتری مهندسی صنایع، گروه مهندسی صنایع، پردیس دانشکده‌های فنی، دانشگاه تهران؛ vahid.khatibi@ut.ac.ir

[2] *One-to-One Marketing*
[3] *Personalization*

ارایهٔ خدمات شخصی‌سازی شده شرکتها را قادر می‌سازد تا نیازها و ارجحیتهای مشتریان را شناسایی کرده و آنها را به مشتریان همیشگی خود مبدّل نموده، بیشترین رضایت و سودبخشی از سوی مشتریان را دنبال کنند. یک روش برای دستیابی به این هدف، پیشنهاد محصولاتی مطابق خواسته‌های مشتری است که از طریق سیستمهای پیشنهادگر[1] تحقق می‌یابد [۴]. سیستمهای پیشنهادگر یکی از رایج‌ترین راه‌حلهای نرم‌افزاری محسوب می‌شوند که در تجارت الکترونیک برای ارایهٔ خدمات شخصی‌سازی شده مورد استفاده قرار می‌گیرند [۵]. این سیستمها با ارایهٔ پیشنهاداتی به مشتریان، مبتنی بر پایهٔ ارجحیتهای آنان و نیز سوابق میلیونها خریدی که روی وبگاه‌های تجارت الکترونیک آنها ثبت شده است، آنها را در یافتن محصولاتی که تمایل به خریدشان دارند کمک می‌کنند [۶].

در این پژوهش به دنبال دستیابی به شیوهٔ کارآمدتری برای سیستمهای پیشنهادگر هستیم تا به کمک آن بخش‌بندی بازار و مشتری را انجام داده و بدین‌ترتیب پیشنهادات بهتری به مشتریان ارایه دهیم. برای حصول این امر، از روشهای داده‌کاوی همچون خوشه‌بندی و قواعد انجمنی[2] بهره برده شده است، به‌طوریکه در فاز اول خوشه‌بندی مشتریان با استفاده از الگوریتم نقشهٔ خودسازمانده (SOM)[3] و نیز الگوریتم K میانگین (K-Means)[4] انجام شده و در فاز دوم با استفاده از قواعد انجمنی برای هر خوشه، نقشه‌ای معتبر انتخاب و بر اساس آن به مشتریان آن خوشه، پیشنهادات مناسب ارائه شده است.

این مقاله به صورت زیر سازماندهی شده است: در بخش دوم به تشریح سیستمهای پیشنهادگر و الگوریتمها و روشهای داده‌کاوی مطرح در این حوزه پرداخته شده است. در بخش سوم، مدل پیشنهادی برای سیستمهای پیشنهادگر ارایه شده و سپس در بخش چهارم نتایج بکارگیری مدل پیشنهادی روی داده‌های یک وبگاه تجاری ایرانی ارایه شده است.

۲. سیستم‌های پیشنهادگر

در این بخش، مفاهیم اساسی دربارهٔ سیستمهای پیشنهادگر و روشهای بکارگرفته شده در آنها همچون خوشه‌بندی و قواعد انجمنی ارایه شده‌اند.

با ظهور و ورود بازاریابی مشتری‌محور بیشتر شرکتها مشتریان را به صورت فردی مورد توجّه قرار داده و به دنبال کسب بیشترین رضایت و سودبخشی برای آنها هستند. یکی از روشهای مورد استفاده برای دستیابی به این هدف، استفاده از سیستمهای پیشنهادگر برای مطابقت محصولات با خواستهٔ مشتریان است [۷]. در سالهای گذشته تعدادی از سیستمهای پیشنهادگر برای کسب‌وکارهای مختلف مورد استفاده قرار گرفته‌اند که از آنها می‌توان به سیستم پیشنهادگر گروپلنز[5] اشاره کرد که کاربران را با پیشنهاد مقالات خبری از وبگاههای خبری یاری می‌رساند. وبگاه رینگو[6] با ارایهٔ پیشنهادات برای موسیقی برخط و نیز فاب[7] با راهنمایی کاربران در قالب صفحات وب و مستندات برخط از این جمله‌اند. همچنین سیستمهای پیشنهادگر بسیاری در زمینهٔ فیلمها و مستندات تصویری، فروشگاه‌های غذایی برخط، وبگاه‌های موسیقی و کتاب‌فروشی‌های برخط همچون آمازون[8] را می‌توان در این حوزه برشمرد.

مشخصه‌ها شاخصهایی هستند که بر اساس آنها دسته‌بندی محصول، بازار و مشتریان انجام می‌پذیرد [۸]. در زمان ایجاد پروندهٔ کاربری[9] مشتریان معمولا از قالبی صریح[10] یا ضمنی[11] از مجموعهٔ داده‌ها استفاده می‌کنند که به آن انتخاب مشخصه‌ها[12] گویند.

از مجموعه داده‌های *صریح* می‌توان به موارد زیر اشاره کرد:

- تقاضا از کاربر برای ارزیابی اقلام در یک مقیاس لغزان
- تقاضا از یک کاربر برای ارزیابی مجموعه‌ای از اقلام بر اساس میزان علاقه‌مندی
- ارائه دو کالا به کاربر و تقاضا برای یک انتخاب بهتر
- تقاضا از کاربر برای ساخت لیستی از اقلام مورد علاقه

از مجموعه داده‌های *ضمنی* می‌توان به موارد زیر اشاره کرد:

- مشاهدهٔ اقلامی که کاربران در فروشگاه برخط از آن بازدید کرده‌اند.
- تجزیهٔ و تحلیل میزان بازدید یک قلم/تعداد کاربر
- نگهداری رکوردهای اقلامی که کاربر در یک خرید برخط تهیه کرده است.
- تجزیه و تحلیل شبکه‌های اجتماعی کاربران و کشف شباهتهای علاقه‌مندی‌ها و دوست‌نداشتنی‌های آنها

با توجّه به اینکه یکی از مهمترین و ضرروی‌ترین عوامل کلیدی موفقیّت در کسب یک بخش‌بندی صحیح و کامل، انتخاب مشخصه‌ها است، لذا در ابتدا به نظریات مختلف در این زمینه توجّه شده است. همانند هر مدلسازی دیگر، انتخاب مشخصه‌هایی

[1] *Recommender Systems*
[2] *Association Rules*
[3] *Self Organizing Map*
[4] *K-Means*
[5] *Group Lens*
[6] *Ringo*
[7] *Fab*
[8] *Amazon.com*
[9] *Profiles*
[10] *Explicit*
[11] *Implicit*
[12] *Feature Selection*

که بتوانند خوشه‌های قابل فهم‌تر ارائه دهند و ارتباط معناداری بین مشخصه‌های هر خوشه کشف کنند، از اهمیّت ویژه‌ای برخوردار است. در این مورد نظریات زیادی وجود داشته، در این قسمت ابتدا روند انتخاب مشخصه‌ها بررسی شده و سپس تعدادی از نظریات در این زمینه معرفی می‌شوند. برخی محققان معتقدند که با گروه‌بندی مشتریان به‌وسیله مشخصات شخصی آنان، می‌توان بهترین پیشگویی را در مورد خرید بعدی آن گروه انجام داد[۹]. اما به تدریج این نظریه که مشتریان با مشخصاتی مانند کلاس اجتماعی و سطح درآمد شبیه به هم، هم‌سلیقه و خرید شبیه به یکدیگر دارند، مورد شک و تردید واقع شد. مشکل دیگری که وجود داشت این بود که برخی از مشتریان علاقه‌ای به گفتن این مشخصات نداشتند و این موضوع باعث می‌شد که نتایج حاصل قابلیّت اطمینان کافی را نداشته باشند و در نتیجه مشخصه‌های رفتار خرید مشتری نیز مورد توجّه قرار گرفت. حال سایر نظریات رایج برای انتخاب مشخصه‌ها بیان می شوند.

۲-۱. نظریهٔ پانج در انتخاب مشخصه‌ها

نظریهٔ پانج[1] مشخصه‌ها را به دو نوع عمومی و خصوصیات محصول محصول و خرید[2] تقسیم می‌کند. مشخصه‌های عمومی شامل سن، سن، جنسیّت، وضعیت تأهل، سطح درآمد، منطقهٔ جغرافیایی، کلاس اجتماعی، سطح تحصیلات، سبک زندگی مشتریان و مانند آن است. مشخصه‌های خصوصیات محصول و خرید شامل سابقهٔ خرید مشتری، نحوهٔ پرداخت پول و نوع محصول درخواستی او است. این نظریه استفاده از ترکیب هر دو نوع مشخصه را پیشنهاد می‌کند، زیرا استفاده از هر کدام این مشخصه‌ها به تنهایی، سازمان را به یک بخش‌بندی جامع از مشتریان خود هدایت نخواهد کرد[۱۰].

۲-۲. نظریهٔ ودل در انتخاب مشخصه ها

نظریهٔ ودل[3] مشخصه‌ها را به چهار بخش مشخصات جغرافیایی، جمعیت‌شناختی[4]، روان‌شناختی و رفتاری مشتریان تقسیم می‌کند. این نظریه کاملاً بر روی مشتری تمرکز کرده و سعی می‌کند یک دید کامل و جامع از مشتری به‌دست آورد. مشخصه‌های جغرافیایی شامل ملیّت، استان، شهر، ناحیه و کشور محل سکونت است. مشخصه های جمعیت‌شناختی شامل سن، جنسیّت، سطح درآمد و تعداد اعضای خانواده است. مشخصه های روان‌شناختی شامل خصلتهای فردی، کلاس اجتماعی و سبک زندگی است. مشخصه‌های رفتاری شامل هدف مورد جستجوی مشتری، دفعات مراجعه، دفعات خرید و وفاداری وی است. این نظریه بر اساس ترکیب این چهار مشخصه عمل می کند[۱۱].

۲-۳. مشخصه های RFM

برای شناخت مشتریان بهره بردن از مشخصه‌های رفتاری از اهمیت ویژه‌ای برخوردار است، اما گاهاً به دلیل در دسترس نبودن آنها از مشخصه‌های جمعیت‌شناختی استفاده می شود. سرنام RFM مخفف سه کلمهٔ Monetary، Frequency و Recency است که به ترتیب معنای "زمان آخرین خرید" (آخرین خرید مشتری چه زمانی بوده است؟)، "تناوب خرید" (مشتری چند وقت یک بار خرید می کند؟) و "پول خرج شده" (مشتری در هر خرید چقدر پول خرج می کند؟) را دارند. بدین‌ترتیب، مشخصه‌های RFM ناظر به بررسی رفتار خرید و عملکرد مشتری است. تعریف دقیق این سه مفهوم عبارت است از:

R: مدت زمانی که از آخرین خرید مشتری می گذرد (که هر چه کمتر باشد، احتمال بازگشت وی و تکرار خرید توسط مشتری مذکور بیشتر است).

F: تعداد خریدهایی که در مدت زمان مشخصی توسط مشتری صورت گرفته است (هرچه تعداد آن بیشتر باشد حاکی از آن است که مشتری موردنظر، در خرید از ما ثبات قدم[5] بیشتری دارد).

M: میزان هزینه‌ای که در مدت زمان مشخصی توسط مشتری، صرف شده است (هرچه مقدار این پارامتر بیشتر باشد، نشان‌دهنده این است که مشتری مورد نظر اهمیّت بیشتری دارد و باید به وی بیشتر توجّه کرد).

پارامترهای مذکور را می‌توان با وزن‌دهی‌های متفاوتی به کار گرفت؛ بسته به اینکه کدام پارامتر اهمیّت بیشتری داشته باشد، می توان وزن آن را در محاسبات بیشتر کرد. به این روش کار WRFM می‌گویند که W در آن مخفّف «وزن‌دهی‌شده»[6] است.

در این روش ترکیبی، همچنین از AHP[7] (فرآیند سلسله مراتبی تحلیلی) نیز استفاده می‌شود که در سال ۱۹۹۴ میلادی توسط ساتی[8] ارائه شد. AHP یک تکنیک ریاضی است که فرد تصمیم‌گیرنده را در فرآیندهای تصمیم‌گیری چندضابطه‌ای[9] یاری می‌کند.

روش ترکیبی مورد نظر ما از ادغام دو روش WRFMمحور[10] و پالایش مشارکتی ترجیحات‌محور[11] استفاده می کند؛ قواعد

---

[1] *Punj*
[2] *Product Specification and Purchase*
[3] *Wedel*
[4] *Demographic*
[5] *Loyalty*
[6] *Weighted*
[7] *Analytic Hierarchy Process*
[8] *Saaty*
[9] *Multi-Criteria*
[10] *WRFM-based*
[11] *Preference-based CF*

انجمنی که از نتایج تحلیل روش WRFM روی داده‌ها، استخراج خواهند شد، برای ارایظ توصیه به مشتریان ثابت‌قدم و باارزش مورد استفاده قرار می‌گیرند و قواعد به‌دست آمده از نتایج روش پالایش مشارکتی ترجیحات‌محور هم برای ارائهٔ پیشنهاداتی به مشتریان سطوح پایین‌تر و کم‌ارزش‌تر بکار گرفته می‌شوند.

می‌توان علاوه بر در نظر گرفتن پیشنهاد برای خریدار، فروشنده را نیز مورد توجّه قرار داد، به‌طوری که با لحاظ کردن سه متغیر احتمال خرید[۱]، احتمال خرید مشتریان مشابه[۲] و سودبخشی محصول[۳] برخی سیستمهای پیشنهادگر را توسعه داده‌اند.

بر این اساس نیز چهار رویکرد تعریف شده است که عبارتند از:

- سیستم پیشنهادگر راحتی-محور (CPRS)[۴]: رویکرد مناسبی است برای فروشندگان جهت پیشنهاد محصولاتی که مکرراً خریداری شده‌اند.
- سیستم پیشنهادگر پالایش مشارکتی-محور (CFRS)[۵]: پیشنهاد محصول بر اساس احتمال خرید بر طبق مشتریان مشابه.
- سیستم پیشنهادگر راحتی و سودآوری-محور (CPPRS)[۶]: پیشنهاد محصول بر اساس احتمال خرید و سودبخشی محصول.
- سیستم پیشنهادگر هیبرید-محور (HPRS)[۷]: پیشنهاد بر اساس سنجش هر دو عامل احتمال خرید مشتریان مشابه و سودبخشی محصول.

احتمال خرید بر دو نوع تعریف می‌شود [۱۲]: احتمال فراوانی-محور[۸] و احتمال شباهت-محور[۹]. این سیستم پیشنهادگر بر اساس اساس الگوهای راهبری[۱۰] و رفتاری[۱۱] مشتری در یک وبگاه عمل می‌کند[۱۳]. الگوهای راهبری عبارتند از: مرور صفحات، کاوش، کلیک محصولات، جایگذاری در سبد و خرید واقعی. همچنین الگوهای رفتاری عبارتند از:

نرخ کلیک برای نوع خاصی از محصول، طول مدت مطالعه صرف شده برای محصولی خاص، تعداد بازدیدهای محصولی خاص، چاپ و نشانه‌گذاری[۱۲].

۳. روشهای داده‌کاوی مورد استفاده در سیستم‌های پیشنهادگر

در این بخش به توصیف روشهای داده‌کاوی رایج در سیستمهای پیشنهادگر پرداخته شده است. بدین منظور ابتدا نقشه‌های خودسازمانده معرفی شده و سپس روش K میانگین و قواعد انجمنی ارایه می‌شوند.

۳-۱. نقشه‌های خود سازمانده

نقشه‌های خودسازمانده (SOM) ابزار قدرتمند و جذّابی برای نمایش داده‌های چندبعدی در فضاهای با ابعاد پایین، معمولاً یک یا دو بعد، فراهم می‌کند. همچنین روشی برای پیش‌پردازش اطلاعات و یا خوشه‌بندی است . نقشه‌های خودسازمانده که گاهی نقشه‌های مشخصهٔ خودسازمانده (SOFM)[۱۳] و یا نقشه‌های کوهونن[۱۴] نامیده می‌شوند، توسط پروفسور کوهونن از آکادمی فنلاند ابداع شد. نقشه‌های خودسازمانده نوعی از شبکه‌های عصبی رقابتی هستند که دارای دو لایهٔ ورودی و خروجی هستند. هر یک از متغیرهای ورودی نرون به یک نرون خروجی متصل است که با یک وزن با هم مرتبطاند. هر نرون خروجی با دیگران در برنده شدن نرون رقابت می‌کند. زمانی SOM قضاوت می‌کند که یک نقطهٔ ورودی به کدام خوشه متعلق است که آن خوشه بتواند آن نرون را برنده کند [۱۴].

۳-۲. الگوریتم K میانگین (K-Means)

این الگوریتم یکی از آسان‌ترین الگوریتمهای یادگیری با سرپرست است که مشکلات بسیاری را در خوشه‌بندی برطرف کرده است و از یک روش آسان و ساده برای طبقه‌بندی مجموعه‌ایی از داده‌ها به تعداد معینی خوشه استفاده می‌کند. ایدهٔ اصلی تعریف K مرکز برای خوشه‌ها است به صورتی که این مراکز تا حد ممکن از هم دور باشند. مرحله بعدی قرار دادن هر داده در خوشهٔ با نزدیکترین مرکز است. در گام بعدی K دوباره محاسبه می‌شود و دوباره داده‌ها در گروه‌های جدید که به مرکز آن نزدیکترند قرار می‌گیرند و این حلقه تکرار می شود و K در هر مرحله تغییر می-کند تا زمانی که دیگر هیچ تغییری صورت نگیرد. هدف این الگوریتم کمینه‌سازی تابع زیر است که به صورت زیر تعریف می-شود :

$$f(x) = \sum_{j-1}^{k} \sum_{i-1}^{n} \left| x_i^j - c_j \right|^2 \qquad (۱)$$

[1] *Purchase probability*
[2] *Purchase probability of similar customers*
[3] *Product profitability*
[4] *Convenience Perspective Recommender System*
[5] *Collaborative Filtering Perspective Recommender System*
[6] *Convenience Plus Profitability Perspective Recommender System System*
[7] *Hybrid Perspective Recommender System*
[8] *Frequency-based probability*
[9] *Similarity-based probability*
[10] *Navigational Patterns*
[11] *Behavioral Patterns*
[12] *Bookmarking*
[13] *Self-Organizing Feature Maps*
[14] *Kohonen Maps*

از آن جایی که $|c^j - c_j|^2$ فاصله بین نقاط داده و مرکز هر خوشه است، یک شاخص محسوب می‌شود [۱۴].

۳-۳. قواعد انجمنی

الگوریتمهای قواعد انجمنی به صورت عمده در یافتن روابط بین اقلام و ویژگیهایی که در پایگاهها موجود است، استفاده می‌شوند تا توانایی درک نقشهایی که احتمالاً در طول خرید در فروشگاهها رخ می‌دهد را داشته باشند [۱۵، ۱۶].

برای نمونه در ۸۰ درصد کسانی که شیر خریداری می‌کنند، لزوماً نان هم خریداری می‌نمایند. لذا تصمیم‌ساز می‌تواند راهبردهای جدید از جمله جابجایی‌های گیشه‌های مرتبط و یا سازماندهی تبلیغات مرتبط را پیاده‌سازی نماید. بنابراین اصلی‌ترین پیشنهاد برای پیاده‌سازی الگوریتم قواعد انجمنی پیاده‌سازی همزمان روابط توسط تجزیه و تحلیل داده‌های تصادفی و استفاده از آنها در نقش مرجع برای تصمیم‌گیری است. قواعد انجمنی به صورت زیر تعریف می‌شوند:

$$I = \{i_1, i_2, \dots, i_m\} \qquad (۲)$$

مجموعه I به عنوان مجموعه‌ای از اعضا است که هر عضو آن یک کالا شناخته می‌شود. D پایگاه کسب‌وکاری است که T بیان کنندهٔ تراکنش هر عضو مجموعه است به شکلی که $\subseteq I$. هر عضو مجموعه که تهی نباشد، زیر مجموعه I است و تنها تشخیص‌دهنده کد TID است. هر عضو مجموعه $X \subset I$ یک مقیاس استاندارد «پشتیبان»[1] برای ارزیابی اهمیت آماری D است، به صورتی که:

$$Support\ (X,D) \qquad (۳)$$

پشتیبان نشان‌دهندهٔ درصد یا تعداد مجموعهٔ تراکنشهایی در D است که شامل هر دوی X و Y $(X \cup Y)$ باشند و مشخص‌کنندهٔ میزان و نرخ کالای X در تراکنش D است.

اطمینان[2] نشان‌دهندهٔ میزان وابستگی یک قلم کالای خاص به دیگری است. در قواعد انجمنی $X \to Y$ که $X,Y \subset I$ و $X \cap Y = \phi$ است. این نقش بدان معنی است که اگر X خریداری شود، بنابراین Y نیز خریداری می‌شود. هر نقش یک مقیاس اندازه گیری استاندارد دارد که اطمینان نامیده می‌شود، به طوری که داریم:

$$Confidence(X \to Y) = \frac{Support(X \cup Y, D)}{Support(X, D)} \qquad (۴)$$

در این مورد $Confidence(X \to Y)$ مشخص می‌کند که اگر کالا شامل X باشد، شانس خرید Y بسیار بالا است. در قواعد انجمنی دو فاز مورد نیاز است. فاز اول ردیابی مجموعه اقلام بزرگ و فاز دوم تولید قواعد انجمنی در استفاده از مجموعه داده‌های بزرگ است. نقشها در دو مورد مورد توجّه قرار می‌گیرند:

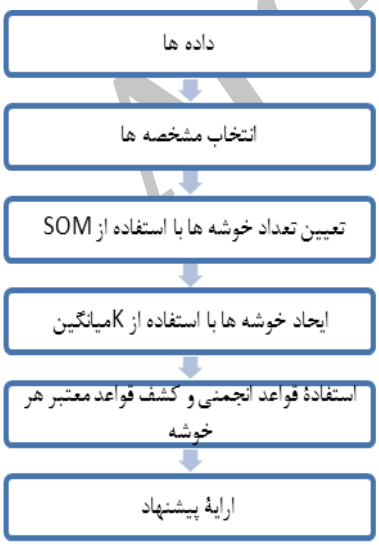
$$Support(X \cup Y, D) \geq Minsup \qquad (۵)$$

$$Confidence(X \to Y) \geq Minconf \qquad (۶)$$

که minsup و minconf هر دو مجموعه‌ای از کاربران هستند [۱۷].

۴. مدل پیشنهادی

در این بخش، مدل پیشنهادی مقاله ارایه می‌گردد. این مدل بر اساس ترکیبی از پالایش محتوا-محور (CB)[3] و پالایش مشارکتی مشارکتی (CF)[4] با تمرکز بر خوشه‌بندی و قواعد انجمنی، پیشنهاددهی به مشتریان انجام می‌دهد. بر اساس مراحلی که در شکل (۱) ارایه شده است، این مدل، پردازش اطلاعات را در دو فاز انجام می‌دهد، به طوری که در فاز اول، پس از انتخاب مشخصه‌ها، تعداد خوشه‌ها با اعمال الگوریتم SOM تعیین شده و سپس جایگذاری عناصر در خوشه‌ها انجام می‌پذیرد. در فاز دوم، با بهره‌گیری از قواعد انجمنی، قواعد حاکم بر هر یک از خوشه‌ها استخراج گردیده و بر اساس آنها به مشتریان، گزینه‌های مناسب پیشنهاد می‌شوند.



شکل ۱. مدل پیشنهادی برای سیستم پیشنهادگر

[1] *Support*
[2] *Confidence*
[3] *Content base Filtering*
[4] *Collaborating Filtering*

در واقع، سیستم پیشنهادگر مبتنی بر این مدل، با توجّه به مشخصه‌های انتخابی و با استفاده از الگوریتم SOM و K میانگین به خوشه‌بندی آنها پرداخته، سپس قواعد معتبر در هر خوشه با استفاده از قواعد انجمنی استخراج می‌شوند. بنابراین بر اساس نقشه‌ای معتبر، پیشنهادت محتمل با تکیه بر نحوهٔ خرید مشتریان در گذشته ارایه می‌گردند. در شکل (۲) روش‌شناسی بکار رفته در توسعهٔ سیستم پیشنهادی ارایه شده است.

بخش ابتدایی این مدل، مشابه سایر مدلهای داده‌کاوی، شامل گردآوری و آماده‌سازی داده‌ها است، به‌طوری‌که داده‌های مورد نیاز از وبگاه ایران‌بین[1] اخذ شده است. این وبگاه فعالیت تجارت الکترونیکی و ارائه خدمات خرده‌فروشی برخط، به ویژه کتابفروشی در فضای مجازی، را بیش از پنج سال است که دنبال نموده است و در حال حاضر یکی از فعّال‌ترین وبگاه‌های تجارت الکترونیک در ایران محسوب می‌شود. داده‌های خام دریافت شده از وبگاه ایران‌بین به صورت چندین فایل جداگانه است که در اولین مرحله تجمیع شده و با توجّه به برنامه‌های نوشته شده توسط پژوهشگر، پاکسازی و نرمال‌سازی داده‌ها برای پردازش انجام گرفت.

شکل ۲. روش‌شناسی توسعهٔ سیستم پیشنهادگر

[1] *IranBin*

۴-۱. انتخاب مشخصه‌ها

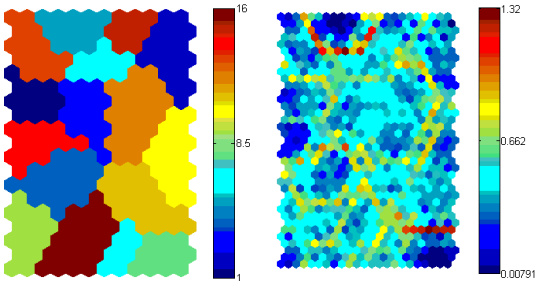
در انتخاب مشخصه‌ها سعی شده است تا از مشخصه‌های جمعیت‌شناختی موجود همچون سن، جنسیت، شغل و تحصیلات استفاده شود، چرا که برای مشخصه‌ها اصولاً شاخصهایی استفاده می‌شوند که بر اساس آنها دسته‌بندی محصول، بازار و مشتری به شکل مناسبی انجام پذیرد. در زمان ساخت پروندهٔ کاربری کاربران، معمولاً از قالبی روشن و یا ضمنی از مجموعهٔ داده‌ها استفاده می‌کنند.

۴-۲. تبدیل داده‌ها به شکل باینری

بعد از انتخاب مشخصه‌ها، برای ارزیابی و انجام محاسبات الگوریتمیک لازم است تا داده‌ها از حالت رشته‌ای مطابق جدول (۱) به حالت باینری تبدیل شوند. بنابراین هر مشخصه با توجّه به ورودی‌هایی که از نظر کاربر به آن اختصاص می‌یابد به بخشهایی تقسیم می‌شود و بر اساس مقدار اختصاص‌یافته، یکی از این بخشها یک و بقیه صفر می‌شوند.

۴-۳. تعیین تعداد خوشه ها

برای تعیین تعداد خوشه‌ها در پایگاه داده مشتریان از الگوریتم بی‌سرپرست SOM استفاده شده است. این الگوریتم پایگاه داده مشتریان را بر اساس مشخصه‌هایی که به شکل باینری درآمده‌اند خوشه‌بندی نموده تا خوشه‌هایی را بر اساس بیشترین تمایز با یکدیگر و نیز وجود عناصر با بیشترین شباهت در هر خوشه تعیین نماید. با توجه به مشخصه‌های خروجی در ماتریسهای U شکل (۳) ترسیم شده است. ماتریس U دارای $2\times Y\text{-}1$ ردیف و $2\times X\text{-}1$ ستون است، در حالی که هر نرون بعدی روی ردیفهای فرد و همه نرونهای روی ردیفهای زوج در ماتریس U، نشان‌دهندهٔ یک فاصله هستند. همچنین ماتریس شباهت، نموداری است که به ما این امکان را می‌دهد تا عناصر مشابه واقع شده در یک نقشهٔ واحد و همچنین مجاورت آنها را با دیگران مشاهده نماییم. اگر بین نمونه‌ها شباهتی نباشد و فاصله بین آنها زیاد باشد، بنابراین واحدهای موجود در ماتریس U به شکل رنگهای گرم نمایش داده خواهد شد.



شکل ۳. ماتریس U و ماتریس شباهت

جدول ۱. تبدیل داده‌ها به شکل باینری

شناسهٔ مشتری	شغل							تحصیلات					
	کارمند	مهندس	مدیر	دانشجو	دانش‌آموز	....	آزاد	لیسانس	فوق لیسانس	فوق دیپلم	زیر دیپلم	دکترا	دیپلم
۳۵۴۲	۱	۰	۰	۰	۰	...	۰	۱	۰	۰	۰	۰	۰
	جنسیت							سن					
	مرد			زن				۵۰ به بالا	۵۰-۴۰	۴۰-۳۰	۳۰-۲۰	۲۰-۱۰	
	۱			۰				۰	۰	۱	۰	۰	

۴-۴. ایجاد خوشه‌ها

بعد از مشخص شدن تعداد خوشه‌ها، برای ایجاد خوشه‌ها از الگوریتم K با سرپرست K میانگین استفاده شده است. بر این اساس، مشتریان با توجّه به مشابهت‌های موجود و نیز بر اساس مشخصه‌های جمعیت‌شناختی و یا رفتار خریدشان در خوشه‌ها قرار می گیرند.

شایان ذکر است برای دلیل بکارگیری روش K میانگین برای خوشه بندی و SOM برای تعیین تعداد خوشه‌ها باید گفت که روش K میانگین یک روش با سرپرست و SOM یک روش بی‌سرپرست است. با استناد به پژوهش [۱۸] که روش خوشه‌بندی دو مرحله‌ای را با ترکیب روشهای SOM و K میانگین برای تحلیل خوشه‌بندی پیشنهاد دادند، در این مقاله، از این الگو بهره برده شده است، چرا که باعث کاهش مشکلات ناشی از فواصل خوشهٔ فازی در SOM می‌شود.

از طرف دیگر پیش‌نیاز روش K میانگین برای خوشه‌بندی آن است که باید تعداد خوشه‌ها را در ورودی در اختیار داشته باشد که این امر با تعیین این عدد از طریق الگوریتم SOM محقق می‌شود. همچنین این روش تاکنون برای سیستم های توصیه‌گر استفاده نشده است.

۴-۵. اعتبار سنجی خوشه‌ها

برای اعتبار سنجی خوشه‌ها روشهای متعارفی وجود دارد. اعتبار سنجی خوشه‌ها در ارزیابی خوشه‌ها از اهمیت بسیار برخوردار است، چرا که بکارگیری نتایج خوشه‌بندی برای کاربردهای مهم نیازمند تأیید ارزیابی است.

در الگوریتمهای مختلف خوشه‌بندی، تعداد خوشه‌ها توسط یک سری پارامتر مقداردهی می‌شوند. رویکردهای زیادی برای یافتن تعداد بهینهٔ خوشه‌ها وجود دارد. از مهمترین شاخصهای اعتبارسنجی می‌توان به این موارد اشاره کرد: دوون[1] ، دیویس-بولدین[2]، سیلوئت[3]، شاخص C، گودمن-کراسکال[4]، انزوا[5]، جک-کارد[6] و شاخص رند[7] [۱۹].

در این تحقیق از روش دیویس بولدین برای اعتبارسنجی خوشه‌ها استفاده شده است. این روش در سال ۱۹۷۶ توسط آقایان دیویس و بولدین ارایه شد، به طوری که آن تابعی است از نسبت مجموع در پراکندگی خوشه به جدایی میان خوشه.

$$\frac{1}{n}\sum_{i=1}^{n} max\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S_n(Q_i,Q_j)}\right\}\frac{1}{n}\sum_{i=1}^{n} max\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S_n(Q_i,Q_j)}\right\} \quad (۷)$$

که در آن N تعداد خوشه‌ها و S_n میانگین فاصلهٔ کلیهٔ اجزا خوشه تا خوشهٔ مرکزی و $S(Q_i,Q_j)$ فاصله بین خوشه‌های مرکزی است که هر چقدر این نسبت کمتر باشد خوشه‌ها متراکم‌تر و از یکدیگر دورتر هستند.

ضمناً در این روش‌شناسی، از خطای عددی[8] و خطای توپوگرافی[9] توپوگرافی[9] نیز استفاده شده است. خطای عددی بر اساس معیار معیار فاصله با مرکز به‌دست می‌آید.

برای محاسبه خطای توپوگرافی، برای هر داده ورودی، اولین و دومین نزدیکترین گره به آن محاسبه شده و اگر این گره‌ها در نقشهٔ خروجی به یکدیگر نزدیک نباشند، خطا اتفاق افتاده است. بنابراین برای اعتبار سنجی، تحلیل و ارزیابی خوشه‌های حاصل به‌وسیلهٔ سه معیار خطای عددی، خطای توپولوژی و دیویس-بولدین محاسبه می‌شود. هر چقدر که این خطاها کمتر باشد، خوشه‌بندی بهتری انجام شده است.

[1] *Duun's Validity Index*
[2] *Davies-Bouldin Validity Index*
[3] *Silhouette Validity Index*
[4] *Goodman-Kruskal Index*
[5] *Isolation Index*
[6] *Jaccard Index*
[7] *Rand Index*
[8] *Quantization Error*
[9] *Topographic Error*

۴-۶. استفاده از قواعد انجمنی

بعد از ایجاد خوشه‌ها، تراکنشهای مشتریان در هر خوشه مشخص شده است. سپس قواعد انجمنی به کمک نرم‌افزار کلمنتاین[1] بر روی مجموعهٔ داده‌های آموزشی با استفاده از الگوریتم اپرایوری[2]، یکی از کارآمدترین روشهای مطرح در قواعد انجمنی، پیاده‌سازی گردید.

بدین‌ترتیب در هر یک از خوشه‌ها به کشف قواعد موجود در آن پرداخته شد و با استفاده از در نظر گرفتن پارامترهای پشتیبان، اعتماد و نظر خبرگان، تعدادی از این نقشها حذف و بقیه لحاظ گردید.

بعد از ایجاد خوشه‌ها و کشف قواعد موجود در هر خوشه امکان ارائه پیشنهاد به مشتری بر اساس خوشه‌ای که در آن قرار می‌گیرد و همچنین مطابق سبد خرید گذشته‌اش مهیا می‌شود.

۴-۷. ارزیابی و ضریب همبستگی

برای ارزیابی، اعتبارسنجی و سنجش ارتباط و همبستگی[3] پیشنهادات ارائه شده توسط سیستم پیشنهادگر در این روش‌شناسی از روش پیرسون[4] استفاده می‌شود. دو متغیر تصادفی x و y را با انحراف معیارهای σx و σy در نظر می‌گیریم. ضریب همبستگی این دو را با ρ نشان داده و به صورت زیر تعریف می‌کنیم:

$$\rho_{x,y} = \frac{\mathrm{cov}(x,y)}{\sqrt{(\mathrm{var}\,x)(\mathrm{var}\,y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad (۸)$$

بدین معنا که داده‌ها به دو دستهٔ داده‌های آموزشی[5] و آزمون[6] تقسیم شده‌اند. در این تقسیم‌بندی تراکنشهای بازهٔ ۲۰۰۷-۲۰۰۸ داده‌های آموزشی و ۲۰۰۸-۲۰۰۹ داده‌های آزمون در نظر گرفته شده‌اند. شایان ذکر اینکه در پژوهشها، مقادیر مختلفی برای ضریب همبستگی در سیستمهای پیشنهادگر ارایه شده است. در حالی که بعضی مقالات، مقادیر ۶۵ درصد به بالا را در نظر گرفته‌اند، در برخی موارد نیز ۳۵ درصد مطرح شده است[۲۰].

۵. موردکاوی مدل پیشنهادی

با توجّه به آنچه در روش‌شناسی مطرح گردید، پیاده‌سازی کلیهٔ مراحل ارایه شده در مدل پیشنهادی در قالب یک موردکاوی در این بخش بررسی می‌شود، به طوری که با پیاده سازی این مدل روی داده‌های وبگاه تجارت الکترونیک ایران‌بین، نتایج حاصل مورد بررسی و ارزیابی قرار می‌گیرند.

با توجّه به مکاتبات صورت گرفته و با همکاری مدیران وبگاه ایران‌بین، این امکان مهیا شد تا پژوهشگران بعد از در اختیار گرفتن داده‌های این شرکت از سال ۲۰۰۷ تاکنون در خصوص کتب فروخته شده توسط این وبگاه، مدل خود را پیاده‌سازی و بیازمایند.

براین اساس در پایگاه مشتریان کتابفروشی، پس از خوشه‌بندی بر اساس مشخصه‌های جمعیت‌شناختی آنها، سیستم پیشنهادگر با استفاده از پایگاه تراکنشها، تراکنشهای انجام شده در هر خوشه را مشخص کرده و به یافتن الگوهای خرید بر اساس قواعد انجمنی در هر خوشه می‌پردازد و بدین‌ترتیب بر اساس فراوانی و موضوع کتاب مورد نظر مشتری بر اساس خریدهای وی، پیشنهادات خود را ارایه می‌نماید.

مشخصات داده‌های موجود در پایگاه داده این وبگاه مطابق جدول (۲) است. در مرحلهٔ آماده‌سازی، داده‌ها پاکسازی و نرمال‌سازی شده‌اند.

جدول ۲. داده های ۲۰۰۹-۲۰۰۷ مشتریان ثبت شده

تعداد تراکنشها	۱۰۳۴۶
تعداد سبد خرید (شماره فاکتور)	۳۰۲۴
تعداد مشتریان دارای خرید	۲۰۶۲
تعداد محصولات خریداری شده	۶۹۰۶
تعداد نوع محتوا	۲۳

در جدول (۳)، انواع گوناگون مشخصه‌های مشتریان و نیز عوامل هر یک از آنها ارایه شده است. همچنین مطابق جدول (۴)، ابتدا در پایگاه مشتری مشخصه‌های جمعیت‌شناختی مانند جنسیت، سن، شغل و تحصیلات، اساس خوشه‌بندی مشتریان قرار داده شده است. با انتخاب این مشخصه‌ها مشتریان بر اساس حوزهٔ تحصیلی و شغلی و نیز دیگر موارد مطرح در خوشه‌های مناسب جای می‌گیرند.

بعد از گردآوری و آماده‌سازی داده‌ها و تعیین مشخصه‌های جمعیت‌شناختی بر اساس مشخصه‌های انتخابی در پایگاه مشتری، داده ها مطابق جدول (۱) به شکل باینری درآمده و تعداد خوشه‌ها توسط الگوریتم SOM مشخص می‌شوند. این خوشه‌بندی بر روی ۲۰۶۲ مشتری انجام پذیرفت که با توجّه به مشخصه‌های انتخابی در ۱۶ خوشه تقسیم شدند.

[1] *Clementine*
[2] *Apriori*
[3] *Correlation*
[4] *Pearson*
[5] *Train Data*
[6] *Test Data*

جدول ۳. انواع گوناگون مشخصه‌های مشتریان و عوامل هر یک از آنها

مشخصه‌های مشتریان		
مشخصه‌های جغرافیایی	*مشخصه‌های جمعیت‌شناختی*	*مشخصه‌های روانشناسی*
• ملیت	• سن	• خصلتهای فردی
• ایالت/استان	• جنسیّت	• کلاس اجتماعی
• شهر	• سطح درآمد	• سبک زندگی
• کشور محل سکونت	• تعداد اعضای خانواده	
	• سطح تحصیلات	
	• شغل	

جدول ۴. مشخصه‌های جمعیت‌شناختی

جنسیت	سن	تحصیلات	شغل	
مرد	۱۰ تا ۲۰	زیر دیپلم	آزاد	گرافیست و نقاش
زن	۲۰ تا ۳۰	دیپلم	بازاریاب	مترجم
	۳۰ تا ۴۰	فوق دیپلم	برنامه‌نویس	مدیر
	۴۰ تا ۵۰	لیسانس	بیکار	معلم، دبیر، مدرس
	۵۰ به بالا	فوق لیسانس	پزشک	مهندس
		دکترا	پژوهشگر	نویسنده
			حسابدار	هیئت علمی
			خبرنگار و روزنامه نگار	وکیل
			دامپزشک	کارگر
			دانش آموز	کارگردان، فیلمساز، فیلمبردار
			دانشجو	کارمند
			طراح	

شایان ذکر اینکه، بر اساس داده‌های موجود و اطلاعات جمعیت-شناختی مشتریان، می‌توان بازنمایی مناسبی از سلایق آنها در خرید محصولات ارائه کرد. همچنین پیش‌پردازش داده‌ها با الگوریتمهای دسته‌بندی و رتبه‌بندی انجام شده است، به طوری که داده‌های مرتب شده این امکان را فراهم ساخت تا دسته بندی‌های واقعی براساس میزان تکرارها در هر یک از شاخصها انجام پذیرد.

در این میان دستورات SQL بکارگرفته شده نقش مهمی در تشخیص الگوهای موجود و تصحیح آنها ایفا نموده است، چرا که فیلدها در ابتدا دارای مقادیر مشخص و از پیش تعریف شده‌ای نبوده‌اند. از طرف دیگر، داده‌هایی که به علّت پراکندگی باعث تشکیل خوشه‌هایی با یک عنصر می‌شدند، با وجود تعداد بسیار کم و ناچیز آنها، با اجماع نظریات مدیران وبگاه حذف شدند.

۵-۱. اعتبارسنجی خوشه‌ها

اعتبار خوشه‌ها توسط معیارهای ارزیابی عددی و توپوگرافیک و همچنین با استفاده از روش دیوید بولدین مطابق جدول (۵) تعیین شده‌اند. با توجه به پارامترهای اشاره شده و ارزیابی اعتبار تعداد خوشه‌ها[1] داریم:

جدول ۵. اعتبارسنجی

خطای عددی	خطای توپوگرافیک	دیویس بولدین
۰/۵۰۴	۰/۰۱۵	۰/۹۰۸۶

همچنان که اشاره شد، برای اعتبار سنجی خوشه‌ها یکی از روشهای مؤثر، دیویس بولدین است که با اعمال آن بعد از اجرای الگوریتم SOM، معمولاً تعداد خوشه را در هر بار پس از اجرا یک عدد تشخیص می‌دهد. برای مثال ممکن است تعداد خوشه‌ها یک بار ۵، بار دیگر ۶ و یا حتی ۸ تشخیص داده شود که برای هر کدام از آنها یک عدد دیویس بولدین، دو خطای عددی و توپوگرافی به‌دست می‌آید. لذا برای تشخیص بهترین تعداد خوشه آن حالتی که عدد دیویس بولدین و دو پارامتر دیگرش کمتر باشد را انتخاب می‌کنیم. مطابق جدول (۶) با استفاده از الگوریتم K میانگین خوشه‌ها ایجاد گردیدند و مشتریان هر خوشه مشخص شدند. با توجه به خروجی‌های متفاوت که در خوشه‌یابی حاصل شد، این مقادیر بهترین حالت تشخیص داده شدند. همچنین خوشه‌ها به وسیلهٔ خبرگان در حوزهٔ کتابفروشی و مدیریت وبگاه تحلیل و ارزیابی شده-اند. نتایج حاصل از ارزیابی خوشه‌ها به قرار جدول (۷) است. شاید این خود در شناخت نیاز بهتر و بیشتر مشتریان، برای مدیران وبگاه کارساز باشد و زمینه‌ساز فراهم‌آوری این نیازها و گسترش بازار در وبگاه مورد نظر باشد. پارامتر نزدیکی و مجاورت[2] به ما

[1] *Cluster Validation*

[2] *Proximity*

کمک خواهد کرد که در فضای چندبعدی بتوان خوشه‌هایی که در مجاورت هر خوشه قرار دارند را تشخیص داد و بر اساس آنها به مشتری پیشنهادات دیگر را نیز مطابق با خوشهٔ مجاورش ارائه نمود.

پس از ایجاد خوشه‌ها و بدست آوردن نقشهای معتبر در هر خوشه، امکان ارایهٔ پیشنهاد به کاربران وجود دارد. برای مثال در در خوشهٔ شماره یک اگر مشتری قبلاً یک کتاب با موضوع "ادبیات" خریداری کرده باشد، لذا سیستم به آن مشتری کتابی با موضوع "روانشناسی" و "علوم اجتماعی" پیشنهاد خواهد کرد جدول (۸). این پیشنهاد بر اساس فراوانی خرید کتابی با موضوع مربوطه خواهد بود.

همچنین اگر مشتری برای اولین بار خریدی انجام دهد تنها بر اساس فراوانی خرید در خوشهٔ مربوطه پیشنهاد انجام می‌پذیرد، به این معنی که در این مثال خاص برای موضوع روانشناسی، کتاب "کیش و مات" و برای موضوع علوم اجتماعی، کتاب "کودک کم‌نوا و خانواده" پیشنهاد داده می‌شوند.

تحلیل حساسیت در روش‌شناسی ارایه شده در بخش تعیین خوشه‌ها اعمال گردید، به طوری که با اعمال تغییراتی روی پارامترهای ورودی تعیین تعداد خوشه‌ها، اثرات آنها روی خروجی‌ها بررسی شد.

با توجّه به اینکه، شاخصهای جامعه‌شناختی دردسترس بر اساس دادگان موجود، از تعداد قابل قبولی برخوردار بوده‌اند، بنابراین تغییر تعداد آنها تاثیر چندانی بر تعداد خوشه‌های مربوطه نداشت، ولی در عین حال در جایگیری عناصر با بیشترین میزان تشابه در خوشه‌ها مؤثر است. لذا با در نظر گرفتن این شرایط، تأثیر تغییرات در گامهای بعدی نیز به حداقل میزان ممکن بوده که نتایج پیاده‌سازی گویای این مطلب است.

جدول ۶. خوشه های ایجاد شده توسط SOM و K میانگین

خوشه ها	تعداد	خوشه‌ها	تعداد	خوشه‌ها	تعداد	خوشه‌ها	تعداد
cluster-1	۲۳۲	cluster-2	۲۴۱	cluster-3	۲۲۲	cluster-4	۴۰
cluster-5	۱۰۴	cluster-5	۱۰۴	cluster-6	۴۳	cluster-7	۲۱۵
cluster-8	۲۶۵	cluster-9	۲۵۰	cluster-10	۵۹	cluster-11	۹۶
cluster-12	۹۳	cluster-13	۶۶	cluster-14	۴۵	cluster-15	۵۱
cluster-16	۴۰						

جدول ۷. تحلیل خوشه‌های ایجاد شده بر اساس مشخصه‌های جمعیت‌شناختی مشتریان

خوشه‌ها	تعداد	سن	جنسیت	تحصیلات	شغل	نزدیکی و مجاورت
خوشه شماره ۱	۲۳۲	۲۰ تا ۳۰	مرد	لیسانس	آزاد	خوشه نه-سه
خوشه شماره ۲	۲۴۱	۱۰ تا ۲۰	مرد	زیردیپلم	دانش‌آموز	خوشه چهارده-سیزده
خوشه شماره ۳	۲۲۲	۳۰ تا ۴۰	مرد	لیسانس	کارمند	خوشه یک-نه
خوشه شماره ۴	۴۰	۳۰ تا ۴۰	زن	لیسانس	کارمند	خوشه پانزده-هشت
خوشه شماره ۵	۱۰۴	۲۰ تا ۳۰	مرد	فوق‌لیسانس	کارمند-دانشجو	خوشه دوازده-ده
خوشه شماره ۶	۴۳	۲۰ تا ۳۰	مرد	دکترا	پزشک-دانشجو	خوشه شانزده-ده
خوشه شماره ۷	۲۱۵	۲۰ تا ۳۰	مرد	دیپلم	آزاد	خوشه سیزده-ده
خوشه شماره ۸	۲۶۵	۲۰ تا ۳۰	زن	لیسانس	کارمند-دانشجو	خوشه پانزده-چهار
خوشه شماره ۹	۲۵۰	۲۰ تا ۳۰	مرد	لیسانس	دانشجو-معلم، دبیر و مدرس	خوشه یک-ده
خوشه شماره ۱۰	۵۹	۲۰ تا ۳۰	مرد	دیپلم	کارمند	خوشه یازده-نه
خوشه شماره ۱۱	۹۶	۲۰ تا ۳۰	مرد	فوق‌دیپلم	بازاریاب-بیکار- معلم، دبیر و مدرس	خوشه ده-یک
خوشه شماره ۱۲	۹۳	۳۰ تا ۴۰	مرد	فوق‌لیسانس	کارمند-هیئت علمی	خوشه پنج-شانزده
خوشه شماره ۱۳	۶۶	۱۰ تا ۲۰	زن	دیپلم	دانش‌آموز-دانشجو	خوشه هفت-چهارده
خوشه شماره ۱۴	۴۵	۱۰ تا ۲۰	مرد	فوق‌دیپلم	دانشجو-بازاریاب	خوشه سیزده-یازده
خوشه شماره ۱۵	۵۱	۱۰ تا ۲۰	زن	دیپلم-لیسانس	آزاد-دانشجو	خوشه چهار-هشت
خوشه شماره ۱۶	۴۰	۳۰ تا ۴۰	مرد	دکترا	دامپزشک-پزشک	خوشه شش-دوازده

جدول ۸. کشف قواعد مبتنی بر سبدهای خرید

خوشهٔ شمارهٔ ۱				
موضوعات خریداری شده	موضوعات پیشنهادی			
ادبیات	روانشناسی	علوم اجتماعی		
علوم انسانی	روانشناسی	علوم اجتماعی	کامپیوتر	علوم مهندسی
علوم اجتماعی	ادبیات	علوم انسانی		
علوم مهندسی	کامپیوتر			

محاسبات انجام شده بر روی داده های وبگاه تجارت الکترونیک ایران‌بین تا انتهای سال ۲۰۰۷ در قالب داده‌های آموزشی است. سپس با استفاده از داده‌های آزمون، ضریب همبستگی برای این مدل مطابق جدول (۹) محاسبه شده است. با توجه به ضریب همبستگی در این روش به نظر می‌رسد که خوشه‌بندی بر اساس مشخصات جمعیت‌شناختی مؤثر واقع گردیده است.

جدول ۹. ضریب همبستگی با استفاده از روش پیرسون

ضریب همبستگی با استفاده از روش پیرسون		
روش اول	داده‌های آزمون	ضریب همبستگی
بر اساس کشف قواعد مبتنی بر سبدهای خرید	۲۰۰۹-۲۰۰۸	۶۸/۲۲۱۵۳

با توجه به نتایج حاصل، این‌گونه به نظر می‌رسد که وبگاه ایران-بین بر روی کتب نظری و کنکوری تمرکز بیشتری دارد چرا که مشتریان آنها بیشترین خرید را در این حوزه انجام داده‌اند. با توجّه به پتانسیل موجود، این وبگاه می‌تواند در ارائهٔ کتب فنی مهندسی و همچنین علوم قرآنی نیز پررنگ‌تر حاضر شود. ضمن اینکه کمترین مخاطبان این فروشگاه را پزشکان تشکیل می‌دهند که این حوزه نیز می تواند مورد توجّه مدیران وبگاه قرار گیرد.

در صورت افزودن فیلدهایی همچون علاقمندیهای مشتریان و نیز رشته تحصیلی در پروندهٔ کاربری آنها نیز می‌تواند امکان پیشنهاد بر اساس محتوا را بهتر مهیا سازد.

بنابراین در تحقیقات آتی می‌توان با استفاده از این ساختارها ارائه یک سیستم پیشنهادگر مبتنی بر روش ترکیبی محتوا-محور و پالایش تجمعی را سامان داد و همچنین می‌توان به جای خوشه-بندی بر روی پایگاه داده مشتریان، بر روی پایگاه داده تراکنشها تمرکز نمود که انتظار می‌رود نتایج بهتری در این روشها کسب شود. ضمن اینکه از روشهای ترکیبی الگوریتم مورچگان با SOM و K میانگین نیز می‌توان برای خوشه‌بندی استفاده نمود.

۶. نتایج

در این مقاله، با بهره‌گیری از روشهای داده‌کاوی نظیر نقشهٔ خودسازمانده و قواعد انجمنی، مدل جدیدی برای سیستمهای پیشنهادگر ارایه گردید که به استفاده از آن می‌توان بخش‌بندی بازار و مشتری را به شیوهٔ کارآمدتری انجام داد و در نتیجه پیشنهادات بهتری به مشتریان ارایه داد. معماری سیستم پیشنهادی در دو فاز به سرانجام رسیده است:

در فاز اول خوشه‌بندی مشتریان براساس مشخصه‌های جمعیت-شناختی سن، جنسیت، شغل و تحصیلات انجام شده که در آن تعداد خوشه‌ها با استفاده از الگوریتم نقشهٔ خودسازمانده (SOM) حاصل شده و سپس خوشه‌ها با الگوریتم Kمیانگین ایجاد گردیده‌اند. در فاز دوم، با استفاده از قواعد انجمنی برای هر خوشه، نقشه‌ای معتبر انتخاب شده و بر اساس آن به کاربرانی که در آن خوشه قرار گرفته‌اند، پیشنهادات مناسب گوناگونی ارائه شده است. مدل مذکور ضمن استفاده از روشهای ترکیبی در خوشه‌بندی و ترکیب آن با مشخصه‌های جمعیت‌شناختی از قواعد انجمنی نیز استفاده نموده است، که از ویژگیهای این مدل محسوب می شود. برای بررسی کارایی مدل پیشنهادی، این سیستم روی داده‌های وبگاه تجاری ایران‌بین اعمال گردید که نتایج حاصل نشان از توان مناسب این روش برای خوشه‌بندی و ارایهٔ پیشنهادات دارند.

مراجع

[1] Campos, L.M.d., J.M. Fernández-Luna, J.F., Huete, *A Collaborative Recommender System Vased on Probabilistic Inference From Fuzzy Observations.* Fuzzy Sets and Systems, 2008. 159(12): pp. 1554-1576.

[2] Chen, L.S., Hsu, F.H., Chen, M.C., Hsu, Y.C., *Developing Recommender Systems with the Consideration of Product Profitability for Sellers.* Information Sciences, 2008. 178: pp. 1032-1048.

[3] Wang, Y.F., Chuang, Y.L., Hsu, M.H., Keh, H.C., *A Personalized Recommender System for the Cosmetic Business.* Expert Systems with Applications, 2004. 26: pp. 427–434.

[4] Kim, K.j., Ahn, H., *A Recommender System using GA K-Means Clustering in an Online Shopping Market.* Expert Systems with Applications, 2008. 34: pp. 1200-1209.

[5] Olmo, F.l.H.n.d., Gaudioso, E., *Evaluation of Recommender Systems: A New Approach.* Expert Systems with Applications, 2008. 35(3): pp. 790-804.

[6] Wang, Y.F., Chuang, Y.L., Hsu, M.H., Keh, H.C., *A Personalized Recommender System for the Cosmetic Business.* Expert Systems with Applications, 2004. 26(3): pp. 427-434.

[7] Al-Shamri, M.Y.H., Bharadwaj, K.K., *Fuzzy-Genetic Approach to Recommender Systems Based on a Novel Hybrid user Model.* Expert Systems with Applications, 2008. 35(3): pp. 1386-1399.

[8] Min, S.H., Han, I., *Detection of the Customer Time-Variant Pattern for Improving Recommender Systems.* Expert Systems with Applications, 2005. 28: pp. 189–199.

[9] Wedel, M., K.W., *Market Segmentation: Conceptual and Methodological Foundations.* Kluwer Academic, Alphen aan den Rijn, Netherlands, 2000.

[10] Punj, S., *Cluster Analysis in Marketing Research: Review and Suggestions for Application.* Journal of Marketing Research, 1983. 20, pp. 134-148.

[11] Chung, K., Han., *Three Representative Market Segmentation Methodologies for Hotel Guest Room Customers.* Tourism Managemeent, 2004. 25(4), pp. 429-441.

[12] Chen, L.S., Hsu, F.H., Chen, M.C., Hsu, Y.C., *Developing Recommender Systems with the Consideration of Product Profitability for Sellers.* Information Sciences, 2008. 178(4): pp. 1032-1048.

[13] Kim, Y.S., Yum, B.J., Song, J., Kim, S.M., *Development of a Recommender System Based on Navigational and Behavioral Patterns of Customers in e-Commerce Sites.* Expert Systems with Applications, 2005. 28(2): pp. 381-393.

[14] Horng-Jinh Chang, L.P.H., Chia-Ling Ho., *An Anticipation Model of Potential Customers' Purchasing Behavior Based on Clustering Analysis and Association Rules Analysis.* Expert Systems with Applications, 2007. 32(3): pp. 753-764.

[15] Changchien, S.W., Lu, T.C., *Mining Association Rules Procedure to Support on-Line Recommendation by Customers and Products Fragmentation.* Expert Systems with Applications, 2001. 20: pp. 325-335.

[16] Nanopoulos, A., Papadopoulos, A.N., Manolopoulos, Y., *Mining Association Rules in Very Large Clustered Domains.* Information Systems, 2007. 32: pp. 649-669.

[17] Kuo, R.J., Lin, S.Y., Shih, C.W., *Mining Association Rules Through Integration of Clustering Analysis and Ant Colony System for Health Insurance Database in Taiwan.* Expert Systems with Applications, 2007. 33: pp. 794-808.

[18] Abidi, S.S., Ong. J., *A Data Mining Strategy for Inductive Data clustering: A Synergy Between Self-Organizing Neural Networks and Kmeans Clustering Techniques*. in Proceedings of *IEEE TENCON 2000*, Kuala Lumpur, Malaysia, 2000, pp. 568-573.

[19] Topchy, A.J., Punch, W., *Combining Multiple Weak Clusterings.* in Proceedings of *The Third IEEE International Conference on Data Mining (ICDM 2003)*, Washington, DC, USA, 2003, pp. 331-338.

[20] Srinivasa Raju, K., Pillai, C.R.S., *Multicriterion Decision Making in River Basin Planning and Development.* European Journal of Operational Research, 1999. 16 (2): pp. 249-257.

[21] Hsu, M.H., *A Personalized English Learning Recommender System for ESL Students.* Expert Systems with Applications, 2008. 34: pp. 683–688.

International Journal of Industrial Engineering & Production Management (2012)

June 2012, Volume 23, Number 1
pp. 1-13

http://IJIEPM.iust.ac.ir/

A Retailing Recommender System Based on Customers' Demographic Features using SOM and Association Rules

Sh. Mosayebian, A. Keramati* & V. Khatibi

Shahab Mosayebian *Information Technology Department, Faculty of Engineering, Tarbiat Modares University, Tehran, Iran*
Abbas Keramati *Industrial Engineering Department, Faculty of Engineering, University of Tehran, Tehran, Iran*
Vahid Khatibi *Industrial Engineering Department, Faculty of Engineering, University of Tehran, Tehran, Iran*

Keywords

Recommender systems,
Online retailing,
Clustering, Self-organizing map,
Association rules

ABSTRACT

The intensive competition in e-Commerce causes effective methods for customer attraction of special importance. In this regard, the recommender systems in commercial websites can precisely determine customers' interests and needs, and offer them most suitable products and services. In this paper, a new model for recommender systems is proposed which segments the market and customers more efficiently, and then provides customers with better offers in two phases; customers are segmented based on demographic features such as age, gender, occupation and education in the first phase. The number of clusters are determined by means of the self-organizing map (SOM), and then clusters are created using K-means. In the second phase, association rules determine a valid map for each cluster which yields the most suitable offers to customers. To examine the efficiency of the proposed model in practice, it is used in an Iranian commercial website, and the results are analyzed.

* **Corresponding author. Abbas Keramati**
Email: *keramati@ut.ac.ir*