علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۸، شماره ۱، ۱۳۹۹
صفحات ۶۵- ۵۶
مقاله عادی

پیاده‌سازی سیستم پیشنهاد شغل ترکیبی مبتنی بر محتوا و پالایش مشارکتی

صدیقه پورنوری[1*]، نرگس پیروی[2]

*نویسنده مسئول، دریافت: ۹۹/۰۶/۰۴، بازنگری: ۹۹/۰۷/۲۵، پذیرش: ۹۹/۰۸/۰۷

[1]دانش‌آموخته کارشناسی ارشد، گروه کامپیوتر، دانشگاه آزاد اسلامی واحد زرقان، زرقان، ایران
[2]مربی،گروه کامپیوتر، دانشگاه آزاد اسلامی واحد زرقان، زرقان، ایران

چکیده

جستجوی شغل دامنه جدید و چالش‌برانگیزی در دسترسی به اطلاعات است. دفاتر کاریابی نیاز به سیستم‌های پیشنهاد دهنده‌ای دارند که بتواند نیازهای کارجویان و فرصت‌های شغلی را پاسخ دهد. در این پژوهش به کمک سیستم پیشنهاد دهنده ترکیبی مبتنی بر محتوا و پالایش مشارکتی و با تحلیل معرفی‌نامه کاربران و اطلاعات فرصت‌های شغلی، مشاغلی متناسب با ویژگی‌های هر فرد، به کاربر توصیه می‌شود. از آن جا که روش پالایش مشارکتی دو مشکل عمده شروع سرد و اعتماد کاربر دارد، در این مقاله علاوه بر حل مشکل شروع سرد، مسئله اعتماد کاربر نیز پوشش داده شده است. روش پیشنهادی شامل سه مرحله خوشه‌بندی اولیه کاربران و تعیین خوشه، تعیین وزن مناسب برای هر ویژگی و تعیین همسایگان نزدیک کاربر جدید، تشکیل ماتریس مجاورتی و محاسبه امتیاز کاربر جدید به هر شغل می‌باشد. جهت خوشه‌بندی اولیه از سیستم محتوا محور بر اساس اطلاعات دموگرافیک کاربران و جهت شناسایی همسایگان نزدیک و ارائه پیشنهادات از پالایش مشارکتی استفاده شده است. جهت انجام آزمایش‌ها از مجموعه داده کاریابی‌های استان فارس استفاده شده و نتایج نشان می‌دهد میزان خطای روش پیشنهادی نسبت به روش پایه سامانه کاریابی ۴,۱٪ کاهش یافته است.

کلمات کلیدی: سیستم پیشنهاد شغل، پالایش مشارکتی، مبتنی بر محتوا، خوشه‌بندی، طبقه‌بندی.

۱- مقدمه

با رشد روزافزون اطلاعات در دنیای حاضر و نیز افزایش چشمگیر کاربران آنلاین[1]، تحویل اطلاعات درست و مناسب به آنها اهمیتی حیاتی یافته است که این امر خود منجر به افزایش تحقیقات در حوزه الگوریتم‌های توصیه‌گر شده است. شرکت‌ها با ساخت سامانه‌های هوشمندی که رفتار گذشته کاربر و نیز رفتار کاربران مشابه را بررسی می‌کنند، پیشنهادات و توصیه‌های مناسب و مورد توجه کاربر را به او ارائه می‌دهند. مواردی از قبیل فرصت‌های شغلی مناسب کاربر، فیلم‌های موردعلاقه کاربر، ویدئوهای پیشنهادی، دوستان فیسبوکی که احتمالاً آنها را می‌شناسید، افرادی که این محصول را خریده‌اند و ... از جمله موارد پیشنهادی به کاربران می‌تواند باشد.

این روزها وب[2] همه چیز را تغییر داده است و شیوه توصیه‌کردن هم از این قضیه مستثنی نبوده است. سیستم‌های پیشنهاددهنده در وب برای توصیه کردن استفاده می‌شوند و به‌عنوان یک رشته علمی درحال‌توسعه می‌باشد، همچنین توانسته کنفرانس‌هایی همچون (ACM) و کارگاه‌های مجزا و مجلات معتبری را به طور انحصاری به خود اختصاص دهد. این سیستم‌ها به کاربران امکان ارائه بازخورد در مورد دوست داشتن یا دوست نداشتن با یک کلیک ساده موس را می‌دهد. ایده اصلی سیستم‌های پیشنهاددهنده، استفاده از منابع مختلف داده‌ها برای پی‌بردن به منافع کاربران می‌باشد. همچنین باید دانست که بدست آوردن اولویت‌های کاربران یک کار حیاتی می‌باشد و سیستم‌های پیشنهاددهنده وابسته به محرکاتی نظیر رفتار انسان، عوامل انسانی، خلق‌وخوی، حواس، احساسات، حالات چهره، ژست بدن و به طور فیزیولوژیکی، تعامل انسان و کامپیوتر می‌باشد.

شایع‌ترین اهداف عملیاتی و فنی سیستم‌های پیشنهاددهنده به‌منظور کسب‌وکار وسیع‌تر و افزایش درآمد شامل مواردی مانند ارتباط (هدیه دادن به کاربران)، تازگی (توصیه اقلامی که کاربر قبلاً ندیده است)، سرندپیتی[3] (خوش‌اقبالی و توصیه موارد غیرمنتظره)، افزایش تنوع توصیه (لیستی از قلم و توجه به حوصله کاربر) می‌باشد. توصیه می‌تواند مستقیم و با صراحت به‌صورت لیستی توصیه‌ای از اقلام و یا توصیه بر اساس اطلاعات توصیفی باشد. همچنین می‌تواند به‌صورت غیرمستقیم از طریق ارتباطات اجتماعی باشد.

۲- سیستم‌های پیشنهاددهنده و انواع الگوریتم‌های مورد استفاده

سیستم‌های توصیه گر از جمله ابزارهایی هستند که می‌توانند کاربران را در محیط‌های الکترونیکی به سمت یافتن اطلاعات، خدمات و کالاهای موردعلاقه هدایت کنند [۱].

سیستم‌های توصیه‌گر (Recommender Systems) به‌منظور شخصی‌سازی اطلاعات پیشنهادی در اواسط دهه ۱۹۹۰ ظهور یافتند. به‌طورکلی وظیفه سیستم‌های توصیه‌گر، بررسی اطلاعات موجود در مورد سلایق کاربران و اولویت‌های آنان، جهت پیش‌بینی علاقه آنها در آینده می‌باشد. نتیجه استفاده از سیستم‌های توصیه‌گر صرفه‌جویی زمانی جستجوهای کاربران است.

۲-۱- هدف سیستم‌های پیشنهاددهنده

هدف سیستم‌های پیشنهاددهنده در واقع رتبه‌بندی آیتم‌های سیستم به لحاظ نزدیک بودن به علایق کاربران می‌باشد تا در هنگام ارائه پیشنهاد، آیتم‌هایی با رتبه بالاتر را به کاربر پیشنهاد دهند. برای این منظور الگوریتم‌های متعددی پیشنهاد شده‌اند که مهم‌ترین آنها عبارتند از:

- Collaborative Filtering (پالایش مشارکتی)
- Content-based Filtering (پالایش مبتنی بر محتوا)
- Social-based Filtering (پالایش مبتنی بر جمعیت‌شناختی)
- Knowledge-based Filtering (پالایش مبتنی بر دانش)
- Context-aware Filtering (پالایش آگاه از زمینه)
- Hybrid Filtering (روش‌های ترکیبی)

۲-۲- روش‌های جمع‌آوری اطلاعات در سیستم‌های پیشنهاددهنده

سیستم‌های پیشنهاددهنده سیستم‌هایی هستند که در پیدا کردن و انتخاب نمودن آیتم‌های مورد نظر کاربران به آنها کمک می‌کنند. طبیعی است که این سیستم‌ها بدون در اختیار داشتن اطلاعات کافی و صحیح در مورد کاربران و آیتم‌های مورد نظر آن‌ها (به‌عنوان‌مثال فیلم، موزیک، کتاب و ...) قادر به پیشنهاد دادن نمی‌باشند؛ بنابراین یکی از اساسی‌ترین اهداف آنها جمع‌آوری اطلاعات گوناگون در رابطه با سلایق کاربران و آیتم‌های موجود در سیستم است. منابع و روش‌های گوناگونی برای جمع‌آوری چنین اطلاعاتی وجود دارد [۲] :

۱. **روش صریح**: روش جمع‌آوری اطلاعات به صورت صریح که در آن کاربر به‌صراحت اعلام می‌کند که به چه چیزهایی علاقه دارد (به‌عنوان نمونه با امتیاز دادن (rate) به یک موزیک).
۲. **روش ضمنی**: این روش اندکی دشوارتر از روش ضمنی است و در آن سیستم باید سلایق کاربر را با کنترل و دنبال‌کردن رفتارها و فعالیت‌های او بیابد. (به‌عنوان‌مثال باید ببینید کاربر به چه موزیک‌هایی بیشتر گوش می‌دهد، چه صفحاتی را بازدید می‌کند، با چه کسانی در ارتباط است و غیره).
۳. **روش استفاده از اطلاعات شخصی کاربران**: علاوه بر اطلاعات ضمنی و صریح برخی از سیستم‌ها نیز هستند که از اطلاعات شخصی کاربران استفاده می‌کنند. به‌عنوان‌مثال سن، جنسیت و ملیت کاربران می‌تواند منبع خوبی برای شناخت کاربر و ارائه پیشنهاد به وی باشد. به این‌گونه از اطلاعات Demographic Information گفته می‌شود که گروهی از سیستم‌های پیشنهاددهنده مبتنی بر همین اطلاعات بنا شده‌اند.
۴. **روش استفاده از اطلاعات موجود در شبکه‌های اجتماعی**: با ظهور web 2 و گسترش شبکه‌های اجتماعی در سال‌های اخیر، محققین به منبع اطلاعاتی دیگری برای بهبود کیفیت پیشنهادات پی بردند که همان اطلاعات موجود در شبکه‌های اجتماعی بود و بر همین اساس کارهای تحقیقاتی زیادی در این حوزه شکل‌ریزی گردید.

۲-۳- اهمیت سیستم‌های پیشنهاددهنده

اهمیت سیستم‌های پیشنهاددهنده وقتی بیش‌ازپیش مشخص می شود که کاربران با انبوهی از اطلاعات و آیتم‌ها مواجه باشند. در چنین شرایطی سیستم باید پیشنهادهای خود را با درنظرگرفتن موارد زیر به کاربر دهد:

۱. شرایط و محیطی که کاربر در آن قرار دارد(user context)
۲. نیاز کاربر
۳. دانش سیستم در مورد کاربر
۴. سابقه فعالیت‌های کاربر

هر کدام از موارد بالا، زمینه ظهور گروهی از سیستم‌های پیشنهاددهنده بوده است. به‌عنوان‌مثال شرایط و محلی که کاربر در آن قرار دارد باعث پیدایش گروهی از سیستم‌های پیشنهاددهنده بنام context-aware recommender system شده و یا دانش سیستم از کاربر، زمینه ظهور سیستم‌های پیشنهاددهنده مبتنی بر دانش (knowledge-based RS) را فراهم آورده است.

۲-۴- نوع داده‌های موجود در بستر سیستم

بنا به کاربرد سیستم، ممکن است انواع مختلفی از منابع اطلاعاتی در سیستم وجود داشته باشد. این اطلاعات می‌توانند شامل موارد ذیل باشند:

- امتیازات کاربران به آیتم‌ها *(ratings)*
- اطلاعات شخصی کاربران
- محتوای مربوط به آیتم‌های سیستم
- ارتباطات موجود در شبکه‌های اجتماعی
- اطلاعات مربوط به موقعیت کار بر *(location-aware information)*

طبیعی است که در پروسه طراحی یک سیستم پیشنهاددهنده باید به نوع داده‌های در اختیار توجه بسیار نمود.

۳- مروری بر تحقیقات پیشین

هدف از این بررسی، ایجاد چشم‌اندازی بهتر در سیستم‌های پیشنهاددهنده و جهت‌گیری آینده‌ای که می‌تواند به شدت بر روی جهان تأثیرگذار باشد، است. در این بخش با بررسی مقالات روز و جنبه‌هایی که نسبت به سایرین اهمیت بیشتری یافته‌اند، اطلاعات جامعی را به‌صورت چکیده، جمع‌آوری کرده‌ایم.

خوشه‌بندی یکی از ابزارهای مهم در سیستم‌های پیشنهاددهنده می‌باشد. الگوریتم‌های خوشه‌بندی به تقسیم داده‌ها به گروه‌های مفید می‌پردازد. به‌طوری‌که مجموعه خوشه‌ها از حداکثر شباهت درون خوشه‌ای و حداقل شباهت برون‌خوشه‌ای برخوردار هستند. الگوریتم خوشه‌بندی CLUTO یک پکیج نرم‌افزاری برای خوشه‌بندی دیتاست‌های کم‌حجم و با ابعاد بالا نیز می‌باشد که برای تجزیه‌وتحلیل ویژگی‌های خوشه‌های مختلف مناسب و کاربردی است [۳].

Chen و همکارانش [۴] با استفاده از خوشه‌بندی‌های تکاملی ناهمگن به جمع‌آوری کاربران با علایق مشابه در خوشه‌های یکسان و سپس کمک به کاربران در یافتن اقلامی که با سلیقه شخصی آنها متناسب است، می‌پردازد. روش آنها دارای دو مرحله است: در مرحله اول، اقلام و کاربران به‌عنوان افراد ناهمگن در شبکه در نظر گرفته می‌شوند. سپس باتوجه‌به مدل شبکه ایجاد شده، حالات افراد در طول زمان تکامل می‌یابند. افراد با امتیاز بالاتر با یکدیگر خوشه‌بندی می‌شوند و همچنین افراد با امتیاز پایین‌تر از هم جدا می‌شوند. پس از تکرار بسیار، حالات اقلام و کاربران با

ثبات و پایدار خواهد شد. در گام بعد باتوجه‌به حالات پایدار ناهمگن افراد، آن‌ها نیز به چند گروه خوشه‌بندی می‌شوند. در مرحله دوم، ورود روش فیلتر مشارکتی مبتنی بر کاربر در هر خوشه است که به واسطه‌ی آن شباهت‌های بین افراد در خوشه یکسان و نه برای همه افراد در سیستم محاسبه می‌شوند و همچنین امتیاز هر خوشه مربوطه، نیز محاسبه می‌شود و در نهایت منجر به توصیه‌های با کیفیت می‌گردد.

باتوجه‌به اینکه پراکندگی و حجم بالای داده‌ها یافتن اقلام مورد نظر را سخت می‌کنند، اما با استفاده از یافتن یک کاربر مجاور و روش خوشه‌بندی فضایی این کار آسان‌تر می‌شود. کوهی و کیانی [۵] باتوجه‌به این موضوع، توانستند روشی را در نظر گیرند که با استفاده از آن زیر فضاهای مختلفی را از اقلام امتیازبندی شده، تحت دسته‌بندی‌های علاقه‌مندی یا عدم علاقه‌مندی استخراج کنند. سپس بر اساس زیر فضاها توانستند ساختارهای درخت را از کاربران مجاور برای کاربر مورد نظر ترسیم کنند.

Hu و همکارانش [۶] باتوجه‌به اهمیت و حساس بودن حوزه‌های بیولوژیکی و تکاملی توانستند با استفاده از سیستم‌های پیشنهاددهنده به پیش‌بینی‌های ژنی بر اساس معیار Gi بیماران بپردازند. در این روش، الگوریتم فیلتر مشارکتی مبتنی بر ژن Gene CF که شامل الگوریتم‌های تکاملی ژنتیک و خوشه‌بندی بر اساس شباهت است، سعی در دسته‌بندی و تشخیص ژن‌های بیماری‌هایی همچون سرطان کبد دارد که این کار به پزشکان در ارائه مراقبت‌های هوشمندانه‌تر و خاص‌تر برای بیماران سرطانی کمک می‌کند.

Lv و همکارانش [۷] برای تولید پیشنهادات معتبر و دقیق، ایده‌ی یکپارچه‌سازی ویژگی‌های معنایی داده‌ها را در هستی‌شناسی فرایند توصیه، پیشنهاد دادند و بدین منظور از بررسی وابستگی بین ویژگی‌ها و پردازش توصیه‌ها از الگوریتم ژنتیک استفاده کردند.

همچنین Ar و Bostasnci [۸] الگوریتم ژنتیک را جهت پالایش مقادیر شباهت کاربر به کاربر در نظر گرفتند که در این رویکرد باتوجه‌به تجزیه‌وتحلیل آماری انجام شده در تعداد انواع همسایگان و معیارهای مختلف شباهت، نتایج امیدوارکننده‌ای را به دست آوردند.

در سال ۲۰۱۶ مقاله‌ای دیگر [۹] با هدف افزایش دقت، ترکیب تکنیک‌ها و کاهش هزینه‌ها مطرح گردید که دارای رویکردی تکاملی، موسوم به کاوش (Invenire) می‌باشد. این روش برای برطرف ساختن تردید در موقعیت‌هایی است که مجبور به انتخاب بین اقلام موجود هستید. این رویکرد روش جدیدی است که از یک الگوریتم جستجو برای بهینه‌سازی تکنیک‌های ترکیبی استفاده می‌نماید و می‌تواند الهام‌بخش تکنیک‌های ترکیبی و سیستم‌های جامع در زمینه‌ی چگونگی انجام اتوماسیون باشد. ازآنجاکه کار کشف یک ترکیب خوب به‌صورت دستی کاری دشوار است، از عملیات ترکیبی به‌صورت خودکار استفاده می‌شود. به همین جهت از الگوریتم ژنتیکی (GA) نیز استفاده می‌کنند. از طرفی رویکرد هوش ازدحامی در سیستم‌های پیشنهاددهنده به‌منظور بهبود عملکرد موتورهای جستجو و کاهش ازدیاد بار اطلاعات با بهره‌برداری از رفتار کاربران جمعی مورد توجه است.

Malizia و همکارانش [۱۰] برای این منظور ایده استفاده از الگوریتم کلونی مورچگان را برای ارائه ویژگی‌هایی جهت دنبال نمودن زمان واقعی اطلاعات ضمنی مشارکتی، در موتورهای جستجو مطرح کردند.

Verma و Katarya [۱۱] با هدف بهبود دقت توصیه برای فیلم‌ها به بهبود عملکرد توصیه و کاهش مشکل مقیاس‌پذیری پرداخته‌اند که برای این کار از خوشه‌بندی K means و الگوریتم تکاملی فاخته استفاده می‌کنند. در این روش، رویکرد بهینه‌سازی و خوشه‌بندی ترکیبی به‌منظور دقت پیش‌بینی فیلم به کار می‌رود. این سیستم برای غلبه بر محدودیت‌های سیستم‌های رایج محتوایی و فیلترینگ مشارکتی استفاده می‌شود. الگوریتم Cuckoo نسبت به سایر الگوریتم‌ها عملکرد بهتری دارد. البته برای رسیدن به بهترین نتایج باید ضریب وزنی مناسب را در بین تمام ضرایب موجود بدست آورد. به همین دلیل از Cuckoo استفاده می‌شود. K means Cuckoo به‌منظور غلبه بر محدودیت‌های سیستم‌های پیشنهاددهنده مشارکتی روش خوشه‌بندی ترکیبی و تکنیکی را برای بهینه‌سازی بهتر جهت دقیق‌تر شدن پیش‌بینی فیلم پیشنهاد می‌کند. نتایج حاصل از این روش به مرور زمان می‌تواند رضایت‌بخش باشد که این مسئله وابسته به مناسب بودن خوشه‌های اولیه است.

امروزه دسته‌بندی بر اساس درجه وقوع مشارکتی اقلام مختلف به شکل جفت‌هایی مشابه در خوشه‌بندی CO کاربردهای زیادی را دارد [۱۲]. اما از طرفی کوهی و کیانی [۱۳] از خوشه‌بندی فازی که نشئت‌گرفته از مدل خوشه‌بندی مشارکتی آماری است، جهت ارائه سیستم پیشنهاددهنده نیز استفاده می‌کنند. خوشه‌بندی فازی به‌عنوان یکی از غالب‌ترین تکنیک‌های خوشه‌بندی مورد استفاده است. یکی از این روش‌های خوشه‌بندی فازی C Means می‌باشد که عملکرد خوبی را داراست. در واقع این روش خوشه‌بندی همان K Means است که با ورود به حوزه فازی تبدیل به C Means شده است.

در فوریه ۲۰۱۶ مقاله‌ای دیگر [۱۴] با هدف پیش‌بینی صفحات وبی که کاربر به آن‌ها مراجعه خواهد کرد، منتشر شد که سبب بهینه‌سازی کیفیت پیشنهاد در صفحات وب گشت. این نوع سیستم‌های پیشنهاددهنده که به اطلاعات متوالی توجه می‌کنند، نوعی سیستم‌های پشتیبانی تصمیم هم هستند که می‌توانند پیشنهاد مطلوبی را در صورت نیاز در اختیار مشتریان قرار دهند. داده‌های وب با فرمت‌های مختلفی از جمله بازدیدهای URL عنوان صفحه وب، هایپرلینک‌های ورودی یا خروجی در صفحه‌ی وب وجود دارند و از خوشه‌بندی‌های فازی جهت ارائه توصیه‌ها استفاده می‌کنند.

در حوزه تجارت الکترونیک از فیلتر مشارکتی چند معیاره نیز می‌توان استفاده کرد و برای بهبود دقت پیش‌بینی در حوزه گردشگری از روش‌های یادگیری ماشین مانند سیستم تطبیق فازی، شبکه عصبی، بردار پشتیبان، رگرسیون به‌عنوان تکنیک‌های پیش‌بینی، تحلیل مؤلفه اصلی به‌عنوان روش کاهش ابعاد، سازماندهی نقشه و حداکثر انتظار به‌عنوان تکنیک خوشه‌بندی استفاده نمود [۱۵].

برای توصیه در وب می‌توان از تکنیک‌های وب‌کاوی و داده‌کاوی باتوجه‌به برخی مشکلات مانند مقیاس‌پذیری، پراکندگی، ارزیابی و شروع سرد استفاده کرد. برای راحتی کار می‌توان از چارچوبی مانند الگوریتم‌های داده‌کاوی باتوجه‌به داده‌های ورودی، ویژگی‌ها و مشخصه‌های کاربران در حوزه‌های خاص استفاده نمود [۱۶].

Bianchi و همکارانش [۱۷] جهت داده‌کاوی بر روی سوابق تماس، چارچوبی را برای شناسایی الگوها و قواعدی که مربوط به یک مشترک عمومی از یک اپراتور تلفن همراه است را در نظر می‌گیرند. در این سیستم، خوشه‌بندی بدون داشتن هیچ دانش پیشینی صورت می‌گیرد. برای داده‌کاوی در این اطلاعات از دو روش بازیابی خوشه‌ها، تجزیه‌وتحلیل ویژگی‌های مشخص جهت تولید فیلد اضافی نیز استفاده می‌شود.

سیستم‌های پیشنهاددهنده مشارکتی نسبت به سایر روش‌ها دارای محبوبیت خاصی در بین محققین می‌باشد اما در برخی مواقع روش فیلتر مشارکتی ضعیف می‌شود. از جمله زمانی که رتبه‌بندی کاربران در اقلام بسیار پراکنده باشد و مشکل پراکندگی داده‌ها در میان باشد.

نجف‌آبادی و همکارانش [۱۸] برای حل این موضوع از سوابق ضمنی کاربران برای اقلام مختلف استفاده کردند تا با به‌کارگیری قوانین وابستگی کاوی بتوانند الگوهای علاقه‌مندی مشابه را کشف نمایند و با استفاده از تکنیک‌های خوشه‌بندی آن‌ها را در گروه‌های مناسبی قرار دهند و در نهایت بتوانند توصیه‌هایی را با کیفیت بالا باتوجه‌به رفع مشکل پراکندگی داده‌ها ارائه دهند.

یکی دیگر از مشکلات سیستم‌های پیشنهاددهنده بحث شروع سرد است و هنگامی عنوان می‌گردد که یک کاربر جدید در سیستم ثبت می‌شود.

Son [۱۹] برای مقابله با این مشکل، ایده استفاده از خوشه‌بندی و الگوریتم‌هایی مانندCS GWFPI M، NHSM، FARAMS، FCF HU را مطرح می‌کند. در واقع شروع سرد زمانی رخ می‌دهد که کاربر جدیدی ثبت می‌شود و این مسئله به دلیل عدم داشتن اطلاعات از وی می‌باشد که منجر به توصیه‌های نامطلوبی می‌گردد و باعث رخداد شروع سرد می‌شود. همچنین از یادگیری عمیق شبکه عصبی برای استخراج ویژگی‌های محتوایی اقلام و پیش‌بینی امتیازبندی‌ها استفاده می‌شود تا شروع سرد به شروع غیرسرد تبدیل گردد [۲۰].

سیستم‌های پیشنهاددهنده به طور اتوماتیک می‌توانند خود را با کاربرانی که علاقه‌مندی‌های متفاوت دارند و همچنین انفجار داده‌ها، به‌سرعت نیز انطباق دهند. رتبه‌بندی شباهت مبتنی بر مجاورت نزدیک‌ترین همسایگی (KNN) یک روش کلاسیک است اما هنوز هم روش محبوب در فیلتر مشارکتی‌هاست. به‌همین دلیل باید روش نزدیک‌ترین همسایگی را بهبود داد.

Luo و همکارانش [۲۱] طرحی را جهت بهبود نزدیک‌ترین همسایگی ارائه دادند که در این روش برای بهبود باید بر روی دو نقطه متمرکز شد:

الف) کاهش پیچیدگی‌های ذخیره‌سازی درحالی‌که دقت پیش‌بینی با استفاده از ضرایب تاس تعمیم‌یافته، حفظ می‌شود.

ب) بهبود دقت پیش‌بینی با یکپارچه‌سازی پشتیبانی شباهت و بایاس‌های خطی و همچنین اجرای افزایشی مربوطه به‌روزرسانی شود.

سیستم‌های پیشنهاددهنده، سیستم‌های کمکی تصمیم‌گیری هستند که رفتار خرید آنلاین پیشین مشتری را تجزیه‌وتحلیل نموده و اطلاعات فعلی محصول را در یافتن اولویت‌های مشتری تطابق می‌دهند و فروشندگان می‌توانند از تکنیک خاص طراحی شده برای تغییر رفتار مصرف‌کننده، استفاده کنند. به همین دلیل به‌دست‌آوردن اولویت‌های کاربر یک کار حیاتی است و نیاز به داده‌کاوی نیز دارد. یکی از راهکارها برای این کار استفاده از روش قانون رابطه مبتنی بر مجموعه ناهموار برای تجزیه‌وتحلیل اولویت‌های مشتری در فرایند تحلیل سلسله مراتبی نیز می‌باشد [۲۲]. حتی برای پیش‌بینی معیارهای عملکرد پست‌های منتشر شده در صفحات برند فیسبوک می‌توان از روش‌های داده‌کاوی استفاده نمود.

یکی از این معیارها "طول عمر پست مصرف‌کنندگان" است که با استفاده از تجزیه‌وتحلیل حساسیت آنها در درک اینکه چگونه هریک از هفت ویژگی ورودی (دسته، لایک‌های کل صفحه، نوع، ماه، ساعت، روز، هفته، پرداخت) می‌توانند توصیه را تحت تأثیر قرار دهند، کاربرد دارد. این روش به همراه درنظرگرفتن نوع محتوا می‌تواند مفید باشد [۲۳].

توصیه‌ها در سیستم‌های پیشنهاددهنده فیلتر مشارکتی بر اساس امتیازدهی سایر کاربران است. اما گاهی این امتیازدهی‌ها به‌تنهایی پاسخگو نیستند و به طور جزئی وارد جوانب مختلف یک قلم نمی‌شوند. از همین رو Yang و همکارانش [۲۴] ایده‌ای را برای دستیابی به درک بهتر رفتار کاربران و توصیه‌های دقیق‌تر مطرح کردند که دارای چارچوبی سه مرحله‌ای است:

۱. کاوش نظرات
۲. محاسبه ضریب
۳. محاسبه وزن جنبه مورد نظر از طریق فاکتورگیری تانسور که جزء تعیین‌کننده امتیاز است.

Verma و Katarya [۲۵] با بیان این موضوع که سیستم‌های پیشنهاددهنده نسبت به رفتار انسان، عوامل انسانی، خلق‌وخوی، حواس، احساسات، حالات چهره، ژست بدن، محرک‌اند، پرداختند و کاربرد این سیستم‌ها را در پیش‌بینی، طبقه‌بندی و مدیریت منسجم عنوان کردند.

در همین راستا Kagita و همکارانش [۲۶] ایده خود را که پیش‌بینی منسجم است، به‌عنوان یک رویکرد نسبتاً جدید برای تعیین کمیت عدم قطعیت در طبقه‌بندی مسائل مطرح کردند و پیاده‌سازی پیش‌بینی منسجم را، وابسته به نوع طبقه‌بندی دانستند. در نهایت یک سیستم پیشنهاددهنده منسجم را طراحی کردند که عدم انطباق را تعریف می‌کند که یک مفهوم کلیدی از پیش‌بینی منسجم، برای سیستم‌های پیشنهاددهنده است. این سیستم نشان می‌دهد که کدام ویژگی‌ها قابلیت معاوضه را دارند یا ندارند.

Fremal و Lecron [۲۷] عقیده داشتند که خوشه‌بندی باتوجه‌به سبک قلم صورت می‌گیرد. همان‌طور که اقلام می‌توانند چندین سبک داشته باشند، ارزیابی‌ها هم می‌توانند در چند خوشه قرار گیرند. هر خوشه پیش‌بینی امتیاز خود و استراتژی‌های وزن‌دهی را بیان می‌کند و سپس برای ترکیب این نتایج از ارزیابی واحدی نیز استفاده می‌شود. معیارهای شباهت هم می‌توانند چندین سبک داشته باشند. معیار شباهت بین اولویت‌ها یک مشکل بسیار مهم برای سیستم پیشنهاددهنده است. می‌توان اولویت‌های مبهم را با استفاده از مجموعه‌های if مدل‌سازی نمود و سپس به استخراج شباهت‌های آنها پرداخت. این روش جهت ایجاد توصیه‌ها و حمایت از تصمیم‌گیری مشتری نیز کاربرد دارد [۲۸].

از دیگر پارامترهای مهم در توصیه‌ها، به هنگام بودن توصیه‌ها می‌باشد. این کار راندمان پیداکردن اقلام را بالا می‌برد. برای این کار Zhang و همکارانش [۲۹] تقسیم‌بندی داده‌های زمانی را در نظر می‌گیرند و لینک‌ها را به مجموعه‌هایی تقسیم می‌کنند. به‌طورکلی نقش زمان‌بر روی لینک‌ها را مدنظر قرار می‌دهند تا با استفاده از این معیار بتوانند توصیه‌هایی با کیفیت بالا را ارائه دهند. اما گاهی کاربران با اقلام و محدوده‌ها آشنا نیستند و سعی در صرفه‌جویی در وقت خود دارند تا در زمان اندکی نیاز خود را برطرف کنند. به همین منظور Scholz و همکارانش [۳۰] بر اساس فرایند پیکربندی محصول توصیه‌هایی را می‌کنند تا کاربران قدرت آشنایی با ویژگی‌های محصول را داشته باشند.

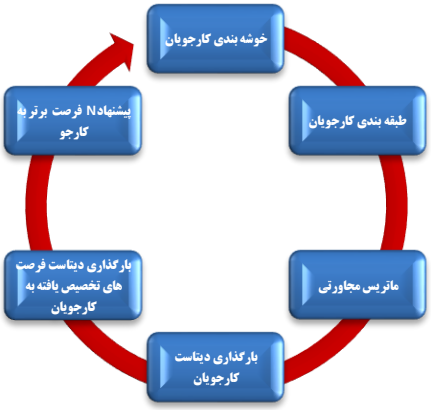
۴- مدل پیشنهادی

روش پالایش مشارکتی یکی از رایج‌ترین روش‌ها در سیستم‌های پیشنهاددهنده است. از نظر مفهومی مبتنی بر یافتن کاربران همفکر یا اقلام مشابه می‌باشد. به دلیل سادگی، شهودی بودن، دقت خوب در پیش‌بینی و تمایل بسیار به بهبود، در این پژوهش مبنای تولید پیشنهادات قرار گرفته است. با وجود موفقیت‌های بسیار روش پالایش مشارکتی مبتنی بر کاربر، این روش با مشکل مقیاس‌پذیری در سیستم‌های با تعداد بالای کاربران و اقلام روبروست و در نتیجه تعداد مقایسه‌های زیادی را برای یافتن شبیه‌ترین کاربران، به سیستم تحمیل می‌کند و سبب تأخیر در ایجاد پیشنهاد در هنگام تعامل با کاربر می‌شود.

این مسئله برای روش‌های مبتنی بر محتوا هم وجود دارد. چرا که با افزایش تعداد کارجویان و شغل‌های سیستم، مقایسه بین تمامی کارجویان (یا مشاغل)، زمان پاسخگویی سیستم را بسیار افزایش می‌دهد. برای حل این مشکل، ترکیب این روش‌ها با روش‌های مبتنی بر مدل پیشنهاد شده است. روش مبتنی بر مدل با تحلیل دوره‌ای داده‌ها به خوشه‌بندی آن‌ها در مدل‌های تخمینی می‌پردازد. از این رو در این پژوهش به دلیل تعداد بالای اقلام (مشاغل تعریف شده توسط کارفرما) و نیز کاربران، در راستای حل مشکل مقیاس‌پذیری، به ترکیب خوشه‌بندی با هر یک از روش‌های پالایش مشارکتی و مبتنی بر محتوا رو آورده شده است.

روش مبتنی بر محتوا به‌تنهایی زمانی قابل استفاده است که ویژگی‌ها را بتوان از اقلام به‌سادگی استخراج کرد و اگر این ویژگی‌ها خیلی گوناگون باشد، نمی‌توان پیشنهادهای خوبی ارائه داد. علاوه بر آن، فقط اقلامی توصیه می‌شوند که دقیقاً همان ویژگی‌ها توسط کاربر مثبت ارزیابی شده باشند که نتیجه آن شخصی‌سازی بیش از اندازه خواهد بود و اقلام گوناگونی توصیه نمی‌شوند. پالایش مشارکتی نیز به‌تنهایی از خلوت بودن ماتریس امتیازات اصلی در ابتدا، رنج می‌برد که همان مسئله شروع سرد است. از این رو برای رسیدن به نتایج بهتر، ترکیبی از روش‌های مزبور

برای بهبود عملکرد و غلبه بر کاستی‌های آن‌ها استفاده شده و برای رسیدن به نتایج بهتر، تکنیک‌های مختلف رویکردهای پالایش مشارکتی و پالایش مبتنی بر محتوا و خوشه‌بندی با یکدیگر ترکیب شده است. با این فرض که پیشنهادات بر اساس شباهت رفتاری و الگوهای عملکردی کاربرانی که شباهت‌های رفتاری و الگوهای مشابهی با کاربر فعلی در گذشته داشته‌اند، ارائه شود.



شکل ۱ - مدل پیشنهادی سیستم توصیه‌گر شغل ترکیبی

در شکل ۱ مدل پیشنهادی سیستم پیشنهاددهنده شغل ترکیبی مبتنی بر محتوا و پالایش مشارکتی مشاهده می‌گردد. همان‌طور که در شکل آمده است در ابتدا برای کاهش حجم محاسبات در داده‌های با حجم بالا از الگوریتم‌های خوشه‌بندی استفاده شده و در مرحله بعد، از طبقه‌بندی داده‌ها برای تقسیم داده‌ها به نسبت ۳۰ به ۷۰ به دو قسمت داده‌های آزمون و آموزش بهره گرفته شده و سپس ماتریس مجاورتی تشکیل می‌شود. در ادامه، مجموعه داده‌های استاندارد شده معرفی‌نامه کارجویان بارگذاری و کارجوی هدف به سیستم وارد می‌شود. پس از آن نیز اطلاعات داده‌های مربوط به مشاغل تخصیص‌یافته به کارجویان در سامانه کاریابی وارد سیستم می‌شود و کارجوی هدف با کارجویان موجود در همان دسته بر اساس وزن‌های اختصاص داده شده به فیلدها مقایسه می‌شود تا شبیه‌ترین Userها به کارجوی هدف مشخص گردند؛ سپس مشاغلی که به کارجوی مشابه پیشنهاد شده به کارجوی هدف نیز پیشنهاد می‌شود.

شرح کامل مراحل انجام پیاده‌سازی مزبور پس از تشریح نحوه‌ی آماده‌سازی داده‌های خام و تبدیل آن‌ها به داده‌های استاندارد بیان می‌گردد. با عنایت به کثرت فراوان داده‌های مورد آزمایش نیاز به خوشه‌بندی و طبقه‌بندی اولیه داده‌ها نیز می‌باشد که نحوه خوشه‌بندی و طبقه‌بندی داده‌ها نیز در جای خود بسط داده خواهد شد.

۴-۱- پاک‌سازی، آماده‌سازی و پیش‌پردازش داده‌های خام

در ابتدای کار، باید داده‌های خام موجود در مجموعه داده‌های اولیه را به داده‌های تمیز و در قالب مورد نیاز سیستم‌های پیشنهاددهنده تبدیل کرد. برای انجام آزمایش‌ها در این مقاله، داده‌های کاریابی‌های استان فارس مورد آزمایش قرار گرفته است. مجموعه داده‌های آزمایش شده شامل ۳ مجموعه داده می‌باشد.

- داده‌های معرفی‌نامه کارجویان
- داده‌های مربوط به اطلاعات فرصت‌های شغلی
- داده‌های فرصت‌های اختصاص‌یافته به کارجویان (یعنی کارجویانی که فرصت‌شغلی به آنها اختصاص یافته است.)

ازآنجاکه داده‌های مستخرج از سامانه کاریابی به‌صورت داده‌های خام در اختیار قرار گرفته است، بایستی قبل از استفاده تبدیل به داده‌های استاندارد گردد. پیش-پردازش یا آماده‌سازی داده‌ها، تأثیر بسیار مهمی بر موفقیت پروژه‌های داده‌کاوی و پیاده‌سازی سیستم‌های پیشنهاددهنده دارد. پاک‌سازی و پیش‌پردازش داده‌های خام نیمی از فرایند ساخت سیستم را به خود اختصاص می‌دهد. فرایند مزبور در شکل ۲ به‌صورت شماتیک آمده است.

شکل ۲ - فرایند تبدیل داده‌های خام به دانش

فرایندی که در فاز پاک سازی، آماده سازی و پیش‌پردازش داده‌های خام اولیه انجام شد، به شرح ذیل می‌باشد:

۴-۱-۱- مراحل پاک‌سازی مجموعه داده اولیه کارجویان

مراحل پاک سازی مجموعه داده‌های اولیه معرفی‌نامه کارجویان به شرح ذیل می‌باشد:

۱. به دلیل محرمانه بودن اطلاعات شخصی کارجویان، ستون‌های مربوط به نام و نام خانوادگی، شماره ملی، تلفن ثابت و همراه و ایمیل، از داده‌های کارجویان حذف گردید.
۲. ستون‌های مربوط به تاریخ تولد و تاریخ فارغ‌التحصیلی به دلیل اینکه در بررسی شباهت کارجویان و فرصت‌های شغلی نقشی نداشتند، حذف شدند.
۳. ستون مربوط به استان به دلیل اینکه کل کارجویان از کارجویانی که در سامانه کاریابی‌های استان فارس ثبت‌نام کرده بودند؛ انتخاب شدند، حذف گردید.
۴. ستون مربوط به سن و وضعیت زبان انگلیسی کارجویان به دلیل اینکه در مجموعه داده‌های اطلاعات فرصت‌های شغلی، معادلی جهت بررسی شباهت نداشت، حذف شد.
۵. ویژگی «مقطع تحصیلی» شامل چهار فیلد می‌باشد که هر کدام معادل یکی از مقاطع تحصیلی (کاردانی، کارشناسی، کارشناسی‌ارشد و دکتری) است.
۶. ویژگی شهر محل سکونت کارجو که مقادیر متنی داشت به مقادیر مناسب عددی تبدیل شد.
۷. باتوجه‌به تعدد فراوان رشته‌های تحصیلی و انواع گرایش‌های مربوط به هر رشته و جهت کاهش حجم محاسبات، رشته‌های تحصیلی، مطابق دفترچه آزمون سراسری دانشگاه‌ها و مراکز آموزش عالی به شاخه‌های تحصیلی تبدیل شد. این نگاشت از ویژگی رشته‌های تحصیلی انجام شد و تعداد ۳۳ شاخه تحصیلی بر اساس رشته‌های تحصیلی کارجویان بدست آمد.
۸. ویژگی شاخه تحصیلی کارجو که مقادیر متنی داشت، به مقادیر مناسب عددی تبدیل گردید.
۹. فیلد «نوع قرارداد» حاوی مقادیر پاره‌وقت، تمام‌وقت و پروژه‌ای بود که کددهی عددی شد.
۱۰. فیلد جنسیت کارجو حاوی مقادیر زن و مرد، کددهی عددی گردید.
۱۱. فیلد وضعیت تأهل حاوی مقادیر مجرد و متأهل نیز کددهی عددی شد.

۴-۱-۲- مراحل پاک‌سازی مجموعه داده‌های اولیه فرصت‌های شغلی

مراحل پاک‌سازی مجموعه داده‌های اولیه فرصت‌های شغلی به شرح ذیل می‌باشد:

۱. ستون عنوان شغل، به دلیل محرمانه بودن از مجموعه داده‌ها حذف شد.
۲. فیلد تاریخ ثبت‌نام و شهر محل کاریابی به دلیل عدم کاربرد حذف گردید.
۳. ویژگی حداقل حقوق به دلیل عدم تطابق مقدار آن در سال‌های مختلف، حذف شد.
۴. ویژگی نوع حقوق بدین دلیل که از نوع فیلد های متنی (غیر قابل دسته‌بندی و گسسته‌سازی) بود، حذف شد.
۵. ویژگی حداقل سابقه کار به دلیل ازدیاد فیلدهای خالی حذف گردید.
۶. در فرایند آماده سازی اطلاعات مشاغل، شهر محل کارگاه که مقادیر متنی داشت به مقادیر مناسب عددی تبدیل شد.
۷. ویژگی رشته تحصیلی مورد نیاز شغل نیز همان گونه که در شرح تبدیل رشته‌های تحصیلی کارجویان به شاخه‌های تحصیلی بیان گردید، تغییر یافت. شاخه‌های تحصیلی مشاغل نیز که مقادیر متنی داشتند به مقادیر مناسب عددی مطابق با کدهای عددی شاخه‌های تحصیلی کارجویان (به استثنای شاخه‌های تحصیلی که در دیتاست کارجویان و فرصت‌های شغلی متفاوت بودند) تبدیل شدند.
۸. ویژگی‌هایی همانند جنسیت، وضعیت تأهل، دارای مقادیری خاص مثل «مرد و زن»، «متأهل و مجرد» و یا مقدار «مهم نیست» بودند. این ویژگی‌ها نیز به مقادیر عددی تبدیل شدند.
۹. ویژگی نوع قرارداد نیز که شامل مقادیر پاره‌وقت، پروژه‌ای و تمام‌وقت بودند به مقادیر عددی تبدیل شدند.
۱۰. ویژگی جنسیت به دو ویژگی زن و مرد تبدیل شد و رکوردهایی که مقدار «مهم نیست» را داشته‌اند، مقدار ۲ می‌گیرد.
۱۱. ویژگی گرایش تحصیلی برای بسیاری از شغل‌ها بدون مقدار بود (همانند شغل‌های مسئول دفتر و یا منشی)؛ بدین معنا که رشته تحصیلی خاصی برای این مشاغل تعریف نشده است؛ بنابراین ویژگی «شاخه تحصیلی آزاد» در کنار شاخه‌های تحصیلی مورد نیاز شغل ایجاد شد و شغل‌هایی که مقدار شاخه تحصیلی آن‌ها بدون انتخاب بود، مقدار آزاد (کد ۱۵) را گرفت، تا مشاغل تخصصی از مشاغل بدون گرایش تخصصی مجزا شوند.
۱۲. «درصد مهم نبودن مقدار ویژگی از نظر کارفرما» شاخصی برای تعیین ارزش و وزن ویژگی در نظر گرفته شده است. لیکن سیستم به‌گونه‌ای طراحی شده است که ضرایب وزن‌ها شناور بوده و باتوجه‌به نظر کاربر قابل تغییر باشد؛ زیرا ممکن است اولویت‌های هر کاربر با کاربر دیگر متفاوت باشد و ملاک کارجویان در انتخاب شغل دلخواه قطعاً یکسان نخواهد بود.

۴-۲- مراحل شبیه‌سازی

سیستم پیشنهاددهنده شغل ترکیبی مبتنی بر محتوا و پالایش مشارکتی موضوع مقاله با مجموعه داده‌های دفاتر کاریابی استان فارس در محیط نرم‌افزاری متلب پیاده‌سازی شده است و شامل مراحل زیر می‌باشد:

۴-۱-۲- خوشه‌بندی کارجویان

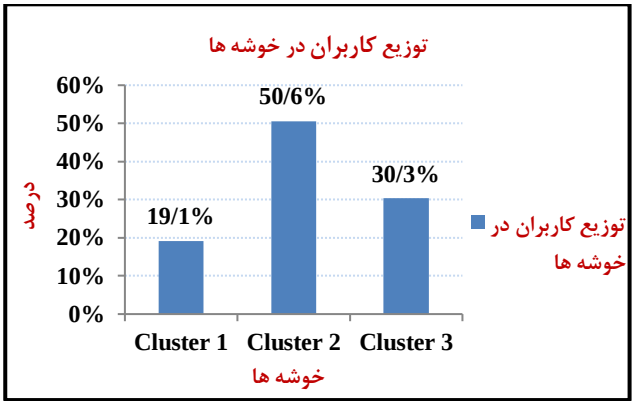
الگوریتم انتخاب شده جهت خوشه‌بندی کارجویان در این شبیه‌سازی، الگوریتم معروف و محبوب K-Means است. در الگوریتم K-Means تعداد خوشه‌ها می‌بایست از قبل تعیین شود؛ به همین دلیل تعداد خوشه متعددی را به الگوریتم خوشه‌بندی K-Means داده و خطاهای مورد نظر را محاسبه نموده و در نهایت تعداد خوشه‌ای انتخاب می‌شود که دارای کمترین میزان خطاست. در این پیاده‌سازی تعداد K را ۳ و تعداد Iteration را ۵ در نظر می‌گیریم. سپس دیتاست کارجویان بارگذاری و فرایند کلاسترینگ روی داده‌ها اعمال می‌گردد و تعداد logاهای هر کلاستر استخراج می‌شود.

نحوه‌ی خوشه‌بندی کارجویان و دقت خوشه‌بندی تأثیر بسیاری در افزایش دقت سیستم خواهد داشت. باتوجه‌به ویژگی شاخه تحصیلی مشاغل، که بیشترین اهمیت را در میان ویژگی‌های شغل دارد، در معرفی‌نامه‌ی کارجویان نیز می‌توان با نگاشت ویژگی‌های تحصیلات کارجو به‌عنوان پاسخی برای نیاز شغل، خوشه‌بندی کارجویان را بر اساس فیلد تحصیلات انجام داد.

جدول ۱ - تعیین تعداد خوشه بهینه با نرم‌افزار وکا

ردیف	K	Itraition	ضریب خطا
۱	۳	۴	۰,۲۹
۲	۴	۴	۰,۲۹
۳	۵	۵	۰,۲۸
۴	۶	۵	۰,۲۸
۵	۷	۶	۰,۲۷
۶	۸	۶	۰,۲۷
۷	۹	۴	۰,۲۴
۸	۱۰	۵	۰,۲۸

همان‌طور که در جدول ۱ مشاهده می‌گردد؛ ضریب خطای محاسبه شده برای Kهای ۳ تا ۱۰ تفاوت قابل ملاحظه‌ای با یکدیگر ندارند؛ بنابراین انتخاب هر کدام از Kهای محاسبه شده در جدول تأثیر چندانی در تغییر نتایج پیاده‌سازی نخواهد داشت. در نهایت با به‌کارگیری روش خوشه‌بندی K-Means، ۳ خوشه‌ی کاربری انتخاب شد. پراکندگی کاربران در خوشه‌ها در شکل ۵ مشاهده می‌شود.



شکل ۳ - توزیع کاربران در خوشه‌ها

۴-۲-۲- طبقه‌بندی کارجویان

در مرحله بعد داده‌ها به نسبت ۳۰ به ۷۰ به دو قسمت داده‌های آزمون و آموزش تقسیم می‌شوند. تعداد سطر و ستون‌ها مشخص می‌شود و فیلدهای اول تا یکی مانده به آخر به‌عنوان ورودی و فیلد آخر به‌عنوان خروجی تعریف می‌شود.

در این مرحله عملیات Classificaition بر روی مجموعه داده‌های استانداردسازی شده اعمال می‌گردد. الگوریتم طبقه‌بندی مورد استفاده در این شبیه‌سازی، الگوریتم درخت تصمیم[۴] می‌باشد. از داده‌های کارجویانی که فرصت‌شغلی به آنان تخصیص‌یافته برای آموزش کارجویان جدید استفاده می‌شود تا بتوان مدلی را تولید نمود که دسته‌ای را برای کاربران جدیدالورود تعیین کند.

۴-۲-۳- ارزیابی متریک‌ها با استفاده از شاخص فولکز- مالوز[۵]

متریک‌ها بر اساس شاخص فولکز – مالوز تعریف شد. برای استفاده از این شاخص باید در مورد چند اصطلاح، آگاهی داشت:

• **مثبت صحیح:** اگر زوجی از مشاهدات که در یک دسته هستند، در یک خوشه نیز قرار بگیرند، نتیجه خوشه‌بندی را برای این زوج «مثبت صحیح» (True Positive) می‌نامیم. مثبت صحیح را با TP نیز نشان می‌دهند.

• **منفی صحیح:** اگر زوجی از مشاهدات که در دو دسته مجزا قرار دارند، در دو خوشه مجزا نیز جای گیرند، نتیجه خوشه‌بندی را برای این زوج «منفی صحیح» (True Negative) می‌نامیم. منفی صحیح را با TN نیز نشان می‌دهند.

• **مثبت کاذب:** اگر زوجی از مشاهدات که در دو دسته مجزا قرار دارند، در یک خوشه جای گیرند، نتیجه خوشه‌بندی برای این زوج «مثبت کاذب» (False Positive) است. مثبت کاذب را با FP نیز نشان می‌دهند.

• **منفی کاذب:** اگر زوجی از مشاهدات که در یک دسته قرار دارند به‌اشتباه در دو خوشه قرار گیرند، نتیجه خوشه‌بندی برای این زوج «منفی کاذب» (False Negative) است. منفی کاذب را با FN نیز نشان می‌دهند.

جدول زیر که وظیفه مقایسه بین تعداد نتایج خوشه‌بندی توسط الگوریتم K-means و دسته‌بندی اصلی (DT) را به عهده دارد، به «ماتریس درهم‌ریختگی»[۶] شهرت دارد.

جدول ۲ - ماتریس درهم‌ریختگی

نتیجه خوشه‌بندی برای زوج (x,y)		دسته‌بندی واقعی	
		در یک دسته	در دو دسته مجزا
خوشه‌بندی	در یک خوشه	TP	FP
	در دو خوشه مجزا	FN	TN

این جدول را به‌اصطلاح ماتریس درهم‌ریختگی می‌گویند. جدول یا ماتریس درهم‌ریختگی، نتایج حاصل از طبقه‌بندی را بر اساس اطلاعات واقعی موجود، نمایش می‌دهد. حال بر اساس این مقادیر می‌توان معیارهای مختلف ارزیابی دسته‌بند و اندازه‌گیری دقت را تعریف کرد. پارامتر دقت(Accuracy)، متداول‌ترین، ساده‌ترین و معیار اندازه‌گیری کیفیت یک دسته‌بند است و عبارت است از میزان تشخیص صحیح دسته‌بند در مجموع دو دسته. این پارامتر در واقع نشانگر میزان الگوهایی است که درست تشخیص داده شده‌اند.

در این شبیه سازی به‌منظور بررسی میزان توانایی سیستم در طبقه‌بندی و تمییز درست اقلام موردعلاقه‌ی کاربر از سه معیار دقت، فراخوانی و معیار F استفاده شده است.

۴-۲-۴- سیستم پیشنهاددهنده

در این مرحله کارجوی هدف انتخاب می‌شود. ویژگی‌های کارجوی هدف از جمله رشته تحصیلی، جنسیت، وضعیت تأهل، شهر محل سکونت، نوع قرارداد و مقطع تحصیلی، بررسی و مشخص می‌شود.

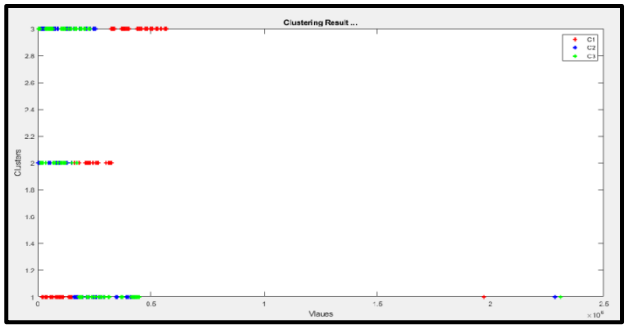
۴-۲-۵- بارگذاری مجموعه داده‌های فرصت‌های تخصیص‌یافته به کارجویان (Dataset Relation) در سیستم

در این مرحله مجموعه داده‌های فرصت‌های تخصیص‌یافته به کارجویان بارگذاری می‌شود. کارجوی هدف با کارجویان موجود در همان دسته بر اساس وزن‌های اختصاص داده شده به فیلدها مقایسه می‌شود تا شبیه‌ترین Userها به کارجوی هدف مشخص گردند. به ویژگی‌هایی که در تعیین تناسب بین فرد و شغل دارای اهمیت بیشتری بوده‌اند، وزن بیشتری اختصاص داده شده است. براین‌اساس به ویژگی‌های شاخه‌ی تحصیلی و شهر که در ثبت شغل‌های سیستم از نظر کارفرما دارای اهمیت بیشتری بوده‌اند، وزن‌های بیشتری اختصاص داده شده است. سپس فرصت‌های اختصاص‌یافته به کارجویان موجود در آن دسته به کارجوی هدف پیشنهاد داده می‌شود. وزن‌دهی پیش‌فرض به ویژگی‌ها درجدول ۳ آمده است.

جدول ۳ - وزن‌دهی به ویژگی‌ها بر اساس پیش‌فرض سیستم

ردیف	ویژگی	وزن
1	شاخه تحصیلی(Branch)	**0.30**
2	شهر(City)	**0.20**
3	جنسیت(Gender)	**0.10**
4	تأهل(Marital)	**0.10**
5	مقطع تحصیلی(Grade)	**0.15**
6	نوع قرارداد(Contract)	**0.15**
	Treshhold : 0-1	

شایان ذکر است وزن‌دهی‌ها شناور بوده و می‌تواند بنابه درخواست و نیاز کارجو و یا کارفرما قابل تغییر باشد. حد آستانه تعریف شده در سیستم نیز مطابق صلاحدید کاربر و بر اساس دقت مورد نظر کاربر سیستم، بین ۰ و ۱ قابل تغییر می‌باشد. بدیهی است هرچه حد آستانه به ۱ نزدیک‌تر باشد، دقت سیستم بیشتر می‌شود. به‌عنوان‌مثال، چنانچه کارجوی جدید مرد و کارجوی مورد بررسی نیز مرد باشد، امتیاز دریافت شده معادل وزن جنسیت (1) و اگر زن باشد، امتیاز دریافتی 0 می‌باشد. در نهایت خروجی سیستم در شکل ۴ نمایش داده می‌شود.



شکل ۴ - نتایج خوشه‌بندی

مساحت زیر نمودار شکل ۴ (Area Under Curve)، به‌عنوان یک معیار برای ارزیابی عملکرد دسته‌بند مورد استفاده قرار می‌گیرد. باتوجه‌به توضیحاتی که پیش‌تر ارائه شد، بدیهی است که در حالت ایده‌آل، مساحت زیر منحنی برابر با بیشترین مقدار خود، یعنی یک است؛ بنابراین، هر چه مساحت زیر نمودار به عدد یک نزدیک‌تر باشد، به معنای بهتر بودن عملکرد دسته‌بند است. محور عمودی نمایش‌دهنده کلاسترها می‌باشد. رنگ قرمز نشانگر خوشه ۱، رنگ آبی نشانگر خوشه ۲ و رنگ سبز نشانگر خوشه ۳ می‌باشد.

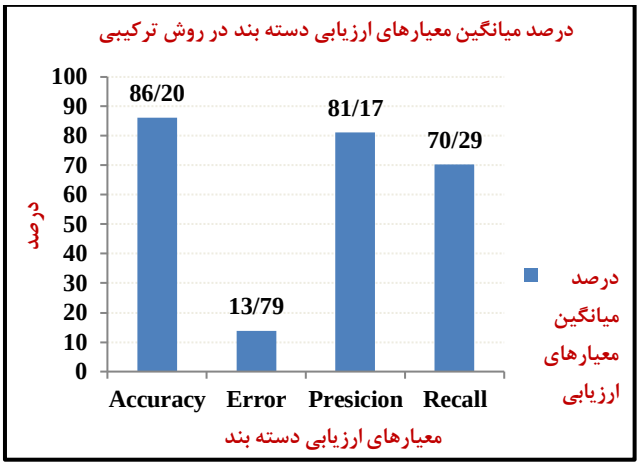
۵- ارزیابی نتایج

خروجی سیستم ترکیبی شامل لیستی از مشاغل پیشنهادی به کارجو می‌باشد که به ویژگی‌های کارجوی هدف بیشترین شباهت را دارد. نمونه‌ای از لیست پیشنهادی سیستم ترکیبی برای کارجو با شناسه کاربری ۱۲۰۰۱۱ در شکل ۵ آمده است. در ابتدای لیست پیشنهاد شغل سیستم ترکیبی، درصد دقت، خطا، پیش‌بینی و بازیابی و پس از آن اوزان ویژگی‌ها و در انتهای لیست نیز فرصت‌های پیشنهادی برای کارجوی هدف نمایش داده می‌شود.

```
Reading Karjoo Dataset...
---------- start Clustering ----------
Dataset(1) Clustered!
---------- start DT ----------
Accuracy(1): 95.6715%
Error(1): 4.3285%
Precision(1): 95.4145%
Recall(1): 91.231%
========== End of File (1) ==========
Selected User ID: 120011
Reading Relation Dataset...
w_Gender = 0.1000
w_marital = 0.1000
w_Branch = 0.3000
w_Grade = 0.1500
w_city = 0.2000
w_contract = 0.1500
treshhold = 0.4500
========== Recommendation List Are ==========
rows_Id = 40 || End Weigth = 1
rows_Id = 44 || End Weigth = 1
rows_Id = 720 || End Weigth = 1
rows_Id = 5392 || End Weigth = 1
rows_Id = 5928 || End Weigth = 1
```

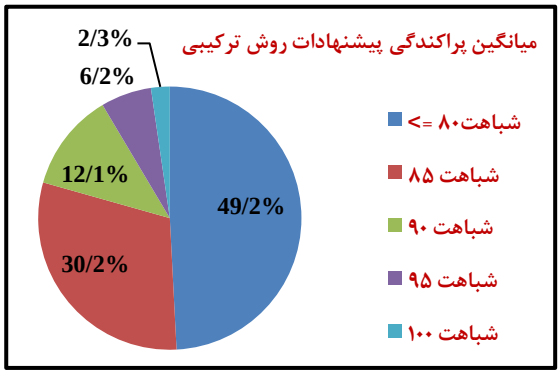
شکل ۵ - خروجی سیستم پیشنهاد شغل به روش ترکیبی

معیارهای دقت، خطا، پیش‌بینی و بازیابی برای گروه نمونه اندازه‌گیری و میانگین هر ۴ معیار موردنظر محاسبه شده است. شکل ۶ میانگین درصد دقت، خطا، پیش‌بینی و بازیابی دسته‌بند در سیستم ترکیبی را نشان می‌دهد.



شکل ۶ - درصد میانگین معیارهای ارزیابی عملکرد دسته‌بند در سیستم پیشنهادی ترکیبی

همان‌طور که در شکل ۶ مشاهده گردید دسته‌بندی سیستم پیشنهادی ترکیبی از دقت نسبتاً بالا و خطای پایین و پیش‌بینی و بازیابی بالایی برخوردار می‌باشد.



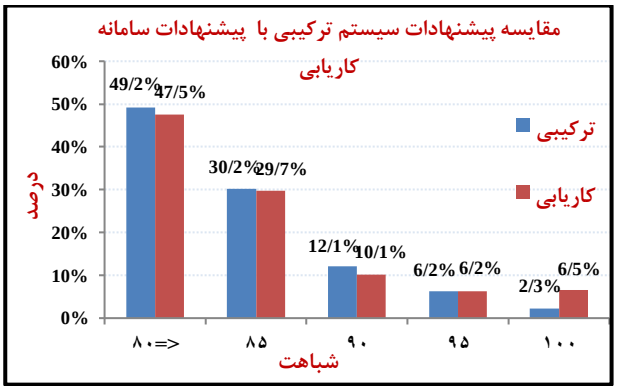
شکل ۷ - میانگین پراکندگی پیشنهادات روش ترکیبی

شکل ۷ میانگین پراکندگی پیشنهادات سیستم پیشنهاد شغل ترکیبی مبتنی بر محتوا، پالایش مشارکتی مبتنی بر کاربر و خوشه‌بندی را نشان می‌دهد.

تحلیل و ارزیابی کلی عملکرد سیستم پیشنهادی به روش ترکیبی، به دلیل تعدد داده‌های پژوهش و افزایش حجم محاسبات تا حدود زیادی زمان‌بر می‌باشد و امکان ارائه آمار دقیق در خصوص درصد صحت و یا خطای پیشنهادات کل سیستم به‌راحتی میسر نخواهد بود. لیکن خروجی سیستم برای تمامی فرصت‌های پیشنهادی به کارجویان، معیارهای ارزیابی سیستم‌های پیشنهاددهنده اعم از دقت، خطا، پیش-بینی و بازیابی را به تفکیک برای هر کارجو محاسبه و در ابتدای لیست پیشنهاد قرار می‌دهد. تنها راه بررسی نتایج کل سیستم، استفاده از آمار استنباطی است. جامعه آماری و نمونه آماری یکی از مباحث اولیه در تحقیق می‌باشد. پژوهشگران معمولاً کار خود را با توصیف اطلاعات شروع نموده (آمار توصیفی) و سعی می‌کنند آن‌چه را از بررسی گروه نمونه به دست آورده‌اند، به گروه‌های مشابه بزرگ‌تر یا جامعه آماری تعمیم دهند. در پژوهش حاضر، بهترین گزینه جهت ارزیابی پیشنهادات سیستم پیشنهادی پژوهش به روش ترکیبی، استفاده از آمار استنباطی است.

برای ارزیابی نتایج حاصل از پژوهش، جامعه آماری شامل کل لیست پیشنهادات فرصت‌شغلی سیستم ترکیبی به کارجویان بوده و بهترین راه‌حل برای کاهش حجم محاسبات، انتخاب یک گروه نمونه می‌باشد. بهترین روش جهت انتخاب گروه نمونه، استفاده از لیست پیشنهادات فرصت‌شغلی به کارجویان در سیستم واقعی کاریابی‌ها (فرصت‌های تخصیص‌یافته به کارجویان در مجموعه داده سوم) می‌باشد که مبنای ارزیابی و تحلیل و مقایسه‌ی نتایج سیستم پیشنهادی پژوهش نیز قرار خواهد گرفت. این مجموعه مشتمل بر تعداد ۴۱۸ رکورد می‌باشد. شناسه‌ی کارجویان مجموعه داده سوم به سیستم پیشنهادی پژوهش به روش ترکیبی وارد و لیست پیشنهادات به کارجویان استخراج و نتایج پیشنهادی سیستم پژوهش با سیستم کاریابی مقایسه شده است.

نتایج حاصل از آزمایشات روش ترکیبی مبتنی بر محتوا، پالایش مشارکتی و مبتنی بر خوشه‌بندی کارجویان موضوع پژوهش که در نرم‌افزار متلب پیاده‌سازی گردید، دقت سیستم در تمامی پیشنهادات را به تفکیک بررسی و برای هر پیشنهاد، دقت پیشنهاد را در ابتدای لیست پیشنهاد قرار داده است. در اکثر موارد دقت پیشنهادات سیستم بالای ۹۰٪ بوده است. لیکن در اندک مواردی که دقت سیستم از ۳۰٪ پایین‌تر آمده ؛ یکی از دلایل اصلی می‌تواند این باشد که برای آموزش داده-های آزمایشی پژوهش از پیشنهادات سامانه کاریابی استفاده شده است و باتوجه‌به اینکه برخی از پیشنهادات سامانه مزبور از شباهت کمتر از ۸۰٪ برخوردار بودند، پیشنهادات سیستم پژوهش به روش ترکیبی نیز در این موارد اندک، بعضاً از دقت کمتر و خطای بالاتری رنج می‌برد. هرچند در مواردی که داده‌های آموزشی از درصد شباهت بیشتری برخوردار بودند؛ پیشنهادات سیستم پژوهشی ترکیبی نیز دقت بالا و خطای پایین‌تری دارند. اما در مجموع صحت پیشنهادات ارائه شده توسط سیستم، توقع نویسندگان را آن‌گونه که انتظار می‌رفت برآورده ساخت.



شکل ۸- مقایسه پیشنهادات سیستم ترکیبی با سامانه کاریابی

شکل ۸ نشان می‌دهد عملکرد سیستم پیشنهاددهنده ترکیبی پژوهش، به لحاظ آماری به سیستم کاریابی بسیار نزدیک می‌باشد که البته چنین آماری قابل پیش‌بینی بود. زیرا داده‌های آزمون پژوهش در روش ترکیبی برای آموزش از داده‌های سامانه کاریابی استفاده کرده‌اند و چون پیشنهادات سیستم ترکیبی بر اساس امتیازدهی کارجویان سیستم کاریابی به فرصت‌ها صورت‌گرفته، چنین اتفاقی کاملاً قابل انتظار بود.

شایان ذکر است روش‌های داده‌کاوی در کنار روش‌های معمول پیشنهاد، سبب بهبود پیشنهادها شدند. شخصی‌سازی پیشنهادات و تناسب ویژگی‌های کارجویان با مشاغل، در امر جستجو مفید واقع شد و علاوه بر آن نیاز کارفرمایان به نیروی متخصص را نیز مرتفع ساخت.

۶- نتیجه‌گیری

در این پژوهش، کارجویان و فرصت‌های شغلی بر اساس ویژگی‌هایی همچون رشته تحصیلی، شهر، جنسیت، وضعیت تأهل، نوع قرارداد، مقطع تحصیلی و دیگر ویژگی‌های آن‌ها دسته‌بندی شدند. سیستم پیشنهاددهنده ارائه شده در این پژوهش توانست بر اساس مشخصات معرفی‌نامه کاربر و فرصت‌های شغلی موجود، فرصت‌های مناسب کاربر را مشخص نموده و صحت بالایی را در پیشنهادها به همراه آورد. همچنین روش‌های داده‌کاوی در کنار روش‌های معمول پیشنهاد، سبب بهبود پیشنهادها شدند. شخصی‌سازی پیشنهادات و تناسب ویژگی‌های کارجویان با مشاغل، در امر جستجو مفید واقع شد و علاوه بر آن نیاز کارفرمایان به نیروی متخصص را نیز مرتفع ساخت.

از بین ویژگی‌های موجود در معرفی‌نامه کارجویان و فرصت‌های شغلی، رشته‌ی تحصیلی، شهر، مقطع تحصیلی، نوع قرارداد، جنسیت و وضعیت تأهل به ترتیب از اهمیت بیشتری در ساخت سیستم برخوردار هستند. هرچند اولویت ویژگی‌های مزبور می‌تواند نسبت به خواست و سلیقه کارجو و کارفرما متغیر باشد؛ بنابراین در ساخت سیستم پیشنهادی سعی گردید از وزن‌دهی شناور به ویژگی‌ها استفاده شود تا بتوان پیشنهادات را بر اساس نیاز و سلیقه کارجو و کارفرما تغییر داد.

جهت بالا بردن سرعت سیستم در ارائه پیشنهاد و بهبود عملکرد سیستم و در راستای پایین آوردن حجم محاسبات در مواردی که با انبوه داده مواجه باشند، روش‌های خوشه‌بندی خواهد توانست مفید فایده قرار گیرد. استفاده از الگوریتم‌های دسته‌بندی نظیر درخت تصمیم نیز می‌تواند برای آموزش داده‌های آزمون مثمر ثمر باشد.

باتوجه‌به اینکه پژوهش حاضر برای اولین‌بار بر روی داده‌های استخراج شده از سامانه کاریابی‌ها انجام شده ؛ لذا امکان ارزیابی نتایج حاصل از این تحقیق با آثار مشابه قبلی وجود نداشته و به‌ناچار نتایج پژوهش با نتایج خود سامانه کاریابی‌ها مقایسه شده است (همان مجموعه داده سوم که مربوط به کارجویانی است که فرصت‌شغلی به آن‌ها اختصاص یافته است). نتایج ارزیابی سیستم پیشنهادی پژوهش در مقایسه با سیستم مورد استفاده در سامانه کاریابی‌ها، از دقت بالاتری برخوردار بوده و درصد بهبود کارایی آن نسبت به سیستم کاریابی‌ها در شکل‌های مزبور به نمایش گذاشته شد. نتایج حاصل از پژوهش با نتایج خود سامانه مزبور مقایسه و نشان می‌دهد مجذور خطای سیستم پیشنهادی ۴٫۱٪ نسبت به سامانه‌های کاریابی کاهش داشته است.

۷- مراجع

[۱] ز. مرادی منش، *تحلیل فنی سیستم‌های توصیه‌گر*، برای اخذ مدرک کارشناسی در رشته مهندسی نرم‌افزار، دانشگاه علم و فرهنگ.

[2] D. H. Park, H. K. Kim, Y. Choi, and J. K. Kim, "*A literature review and classification of recommender systems research*," Expert Systems with Applications, vol. 39, no. 11, pp. 10059–10072, 2012.

[3] G. Karypis, and Book: "*CLUTO A Clustering Toolkit*", November 28, 2003.

[4] J. Chen, Uliji, H. Wang, and Z. Yan, "*Evolutionary heterogeneous clustering for rating prediction based on user collaborative filtering*", Swarm and Evolutionary Computation, 2017.

[5] H. Koohi, and K. Kiani, "*A New Method to find Neighbor Users that Improves the Performance of Collaborative Filtering", Expert Systems With Applications*, 2017.

[6] J. Hu, S. Sharma, Z. Gao, and V. Chang, " *Gene-based Collaborative Filtering using recommender system*", Elsevier, Computers and Electrical Engineering, 2017.

[7] G. Lv, Ch. Hu, and Sh. Chen, "*Research on recommender system based on ontology and genetic algorithm*", Neurocomputing, Elsevier, 2015.

[8] Y. Ar, and E. Bostanci, "*A Genetic Algorithm Solution to the Collaborative Filtering Problem*", Expert Systems with Applications, 2016.

[9] E.Q. Silva, C.G. Camilo-Junior, L.M.L. Pascoal, and T.C. Rosa, "*An evolutionary approach for combining results of recommender systems techniques based on collaborative filtering*", Expert Systems With Applications, Elsevier, 2016.

[10] A. Malizia, K.A. Olsen, T. Turchi, and P. Crescenzi, "*An ant-colony based approach for real-time implicit collaborative information seeking*", Information Processing and Management, Elsevier, 2017.

[11] R. Katarya, and O. Verma, "*An effective collaborative movie recommender system with cuckoo search*", Egyptian Informatics Journal, Elsevier, 2016.

[12] K. Honda, "*Fuzzy Co-Clustering and Application to Collaborative Filtering*", Springer International Publishing AG, 2016.

[13] H. Koohi, and K. Kiani, "*User Based Collaborative Filtering using Fuzzy C-Means*", Measurement, 2016.

[14] R. Katarya, and O.P. Verma, " *An effective web page recommender system with fuzzy c-mean clustering*", Multimed Tools Appl Journal in Springer, vol. 9978, 2016.

[15] M. Nilashi, K. Bagheri Fard, M. Rahmani, and V. Rafe, "*A Recommender System for Tourism Industry Using Cluster Ensemble and Prediction Machine Learning Techniques*", Computers & Industrial Engineering, 2017.

[16] M.N. Moreno, S. Segrera, V.F. López, M.D. Muñoz, and A.L. Sánchez, "*Web mining based framework for solving usual problems in recommender systems. A case study for movies' recommendation*", Neurocomputing, Elsevier, 2015.

[17] F. Bianchi, A. Rizzi, A. Sadeghian, and C. Moiso, "*Identifying user habits through data mining on call data records*", Engineering Applications of Artificial Intelligence, Elsevier, 2016.

[18] M. Najafabadi, M. Mahrin, S. Chuprat, and H. Sarkan, "*Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data*", Computers in Human Behavior, Elsevier, 2017.

[19] L.H. Son, "*Dealing with the new user cold-start problem in recommender systems:A comparative review*", InformationSystems, Elsevier, 2017.

[20] J. Wei, J. He, K. Chen, Y. Zhou, and Z. Tang, "*Collaborative Filtering and Deep Learning Based Recommendation System For Cold Start Items*", Expert Systems With Applications, 2016.

[21] X. Luo, Y. Xia, Q. Zhu, and Y. Li, "*Boosting the K-Nearest-Neighborhood based incremental collaborative filtering*", Knowledge-Based Systems, Elsevier, 2013.

[22] Sh. Liao, and H. Chang, "*A rough set-based association rule approach for a recommendation system for online consumers*", Information Processing and Management, Elsevier, 2016.

[23] S. Moro, P. Rita, and B. Vala, "*Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach*", Journal of Business Research, Elsevier, 2016.

[24] C. Yang, X. Yu, Y. Liu, Y. Nie, and Y. Wang, "*Collaborative filtering with weighted opinion aspects*",Neurocomputing, Elsevier, 2016.

[25] R. Katarya, and O. Verma, "*Recent developments in affective recommender systems*", Physica A, Elsevier, 2016.

[26] V. Kagita, A.K. Pujari, V. Padmanabhan, S. Sahu, and V. Kumar, "*Conformal Recommender System*",Information Sciences, 2016.

[27] S. Frémal, and F. Lecron, "*Weighting strategies for a recommender system using item clustering based on genres*", Expert Systems With Applications, Elsevier, 2017.

[28] P. Ładyzynski, and P. Grzegorzewski, "*Vague preferences in recommender systems*", Expert SystemsWithApplications, Elsevier, 2015.

[29] F. Zhang, Q. Liu, and A. Zeng,"*Timeliness in recommender systems*", Expert Systems with Applications, Elsevier, 2017.

[30] M. Scholz, V. Dorner, G. Schryen, and A. Benlian, "*A configuration-based recommender system for supporting e-commerce decisions*", European Journal of Operational Research, 2016.

صدیقه پورنوری کارشناس ارشد خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه آزاد اسلامی واحد زرقان اخذ نموده است. زمینه پژوهشی موردعلاقه نامبرده سیستم‌های پیشنهاد دهنده (Recommender Systems) است.

آدرس پست الکترونیکی ایشان عبارت است:

sedighehpoornouri@yahoo.com

نرگس پیروی دانشجوی دکتری در رشته مهندسی فناوری اطلاعات – گرایش تجارت الکترونیک است. ایشان عضو هیات علمی تمام وقت (رسمی قطعی) و مدیر گروه کامپیوتر دانشگاه آزاد اسلامی واحد زرقان می‌باشد. زمینه پژوهشی موردعلاقه نامبرده داده های حجیم (Big Data) است.

آدرس پست الکترونیکی ایشان عبارت است:

n.peyravi@zariau.ac.ir

[1] online
[2] World Wide Web
[3] serendipity

[4] Decision Tree
[5] Fowlkes-Mallows
[6] Confusion Matrix

The CSI Journal on Computing Science and Information Technology

The CSI Journal on
Computing Science and
Information Technology
Vol. 18, No.1, 2020

Implementation Hybrid job recommender system Content-based and Collaborative filtering

Sedigheh Poornouri [1], Narges Peyravi [2]

[1,2] Department of Computer, Islamic Azad University Zarghan Branch, Zarghan, Iran

Abstract

Job search is a new, important and challenging domain in accessing information. Employment agencies need recommender systems that can meet the needs of job seekers and job opportunities. In this research, with the help of a content-proposing system based on content and collaborative refinement, and by analyzing users' referrals and job opportunities information, jobs tailored to each individual's characteristics are recommended to the user. Since the participatory filtering method has two main problems of cold start and user trust, in this article, in addition to solving the cold start problem, the issue of user trust is also covered. The proposed method includes three stages of initial clustering of users and clustering, determining the appropriate weight for each feature and determining the new user's close neighbors, forming an adjacent matrix and calculating the new user score for each job. For initial clustering, content-based system based on users' demographic information and for identifying close neighbors and making suggestions, participatory refinement has been used. To perform the experiments, the job search data set of Fars province was used and the results show that the error rate of the proposed method has been reduced by 4.1% compared to the basic method of the job search system.

Keywords: Job recommender system, Collaborative filtering, content-based, clustering, classification