خوشه‌بندی کاربران داده‌های دریایی با استفاده از تکنیک داده‌کاوی

فرناز غیاثی[1] | **نوید نظافتی**[2] | **سجاد شکوهیار**[3]

۱. [پدیدآور رابط] کارشناسی ارشد مدیریت فناوری اطلاعات؛ دانشگاه شهید بهشتی؛ سرپرست مرکز داده‌ها؛ پژوهشگاه ملی اقیانوس‌شناسی و علوم جوی farpar2002@yahoo.com

۲. دکتری مهندسی صنایع؛ استادیار؛ دانشگاه شهید بهشتی nezafati@sbu.ac.ir

۳. دکتری مهندسی صنایع؛ استادیار؛ دانشگاه شهید بهشتی s_shokouhyar@sbu.ac.ir

مقالهٔ پژوهشی

دریافت: ۱۳۹۳/۰۹/۱۹
پذیرش: ۱۳۹۳/۱۱/۱۳

دورهٔ ۳۰ شماره ۴
ص‌ص. ۱۰۲۵-۱۰۴۹

چکیده: هدف از این تحقیق خوشه‌بندی کاربران داده‌های دریایی با استفاده از تکنیک داده‌کاوی است. با محقق‌شدن این هدف، سازمان‌های دریایی قادر به شناخت داده‌های موجود و همچنین اطلاع از نیازهای کاربران خود خواهند شد. در این تحقیق برای پیاده‌سازی داده‌کاوی از مدل استاندارد CRISP-DM استفاده شده است. داده‌های مورد نیاز از اطلاعات و پروفایل ۵۰۰ کاربر داده‌های دریایی از سال ۱۳۸۶ تا ۱۳۹۳ در پژوهشگاه ملی اقیانوس‌شناسی و علوم جوی استخراج شده است. برای خوشه‌بندی از الگوریتم TwoStep استفاده شده است. در این تحقیق، برای نخستین بار با استفاده از خوشه‌بندی، الگویی میان کاربران داده‌های دریایی اعم از دانشجو، سازمان و پژوهشگر، و اطلاعات داده‌های مورد درخواست آنها (منبع داده، نوع داده، مجموعهٔ داده، پارامتر و منطقهٔ جغرافیایی) کشف شد. مهم‌ترین خوشه‌های به‌دست آمده عبارت‌اند از کاربر دانشجو با منبع دادهٔ بین‌المللی، کاربر دانشجو با نوع دادهٔ شیمی دریا، کاربر دانشجو با مجموعه دادهٔ «پایگاه دادهٔ اقیانوسی جهانی»، کاربر سازمان با پارامتر نیترات و کاربر دانشجو با منطقهٔ جغرافیایی خلیج فارس. کشف این الگوها، مدیران ارشد را قادر می‌سازد تا به‌درستی در مورد داده‌های موجود خود و برنامه‌ریزی برای جمع‌آوری داده در آینده تصمیم‌گیری کنند و درک بهتری از نیازهای کاربران خود داشته باشند و کاربران داده نیز در راستای تقاضای خود هدایت شوند. در پایان پیشنهاداتی به‌منظور بهبود عملکرد سازمان‌های دریایی ارائه شده است.

کلیدواژه‌ها: استاندارد CRISP-DM؛ الگوریتم TwoStep؛ خوشه‌بندی؛ داده‌کاوی؛ دادهٔ دریایی

پژوهشنامهٔ پردازش و مدیریت اطلاعات
فصلنامه | علمی پژوهشی
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISA، ISC و Scopus
http://jipm.irandoc.ac.ir
پژوهشگاه علوم و فناوری اطلاعات ایران

۱. مقدمه

سازمان‌های تحقیقاتی دریایی داخل کشور برای انجام کارهای پژوهشی خود و ارائهٔ خدمات به سایر ارگان‌ها، به جمع‌آوری داده‌های دریایی از سازمان‌های داخلی و خارجی، ماهواره‌ها و گشت‌های علمی در محیط‌های آبی ایران (خلیج فارس، دریای عمان و دریای خزر) می‌پردازند. این داده‌ها از حجم زیادی برخوردارند و انواع مختلفی از مجموعهٔ داده‌ها، انواع داده‌های دریایی از جمله فیزیک دریا، شیمی دریا، بیولوژی دریا ... و همچنین پارامترها را شامل می‌شوند.

با توجه به اهمیت حوضه‌های آبی و نیاز روزافزون به ثبت داده‌های دریایی به نظر می‌رسد امروزه جامعهٔ دریایی کشور با پایگاه‌های حجیمی از داده‌های دریایی روبه‌روست (وفایی، چگینی، سقایی، و عظام ۱۳۹۲). دستگاه‌ها و تجهیزات جمع‌آوری این داده‌ها گران و نگهداری و مراقبت از آنها بسیار پرهزینه است. همچنین برگزاری گشت‌های دریایی نیازمند صرف بودجهٔ فراوان می‌باشد. بنابراین، استفاده و به‌کارگیری از این داده‌ها و اطلاعات حاصل از پردازش این داده‌ها، حائز اهمیت است. به همین دلیل، نیاز به ابزاری می‌باشد که بتوان داده‌های ذخیره‌شده را پردازش نمود و اطلاعات حاصل از این پردازش و دانش کشف‌شده را در اختیار کاربران قرار داد.

ابزار داده‌کاوی کمک می‌کند تا خدمات، متناسب با نیاز کاربران ارائه شوند و بر اساس اهمیت تقاضاها، استراتژی لازم جهت ارائهٔ بهتر خدمات در پیش گرفته شود. داده‌کاوی به بهره‌گیری از ابزارهای تجزیه و تحلیل داده‌ها به‌منظور کشف الگوها و روابط معتبری که تاکنون ناشناخته بوده‌اند، اطلاق می‌شود (وفایی، چگینی، سقایی، و عظام ۱۳۹۲).

داده‌کاوی، سازمان‌ها را قادر می‌سازد تا از سرمایهٔ داده‌هایشان به‌درستی بهره‌برداری نمایند و از این ابزار برای پشتیبانی فرایند تصمیم‌گیری استفاده کنند. داده‌کاوی، پردازش بهینهٔ تصمیم‌گیری را در سازمان‌ها تسهیل می‌کند و از طریق استخراج دانش باارزش از داده‌ها، تصمیم‌گیری را برای مدیران سازمان‌ها تسهیل می‌کند. بنابراین، ضروری است که برای استفاده از این ابزار در سازمان‌ها اهمیت بیشتری قائل شد تا در نهایت به فرایند تصمیم‌گیری بهینه برای مدیران منجر شود.

هدف اصلی این تحقیق، خوشه‌بندی کاربران داده‌های دریایی بر اساس ویژگی‌های

دادهٔ درخواستی مانند نوع داده[1]، مجموعهٔ داده[2]، پارامترها[3]، منبع داده[4] و منطقهٔ جغرافیایی با استفاده از تکنیک داده‌کاوی است.

در این تحقیق قصد داریم به سؤالات زیر پاسخ دهیم:

سؤال اصلی: چه الگویی بین کاربران داده با نوع داده، مجموعهٔ داده، پارامترهای داده، منبع داده و منطقهٔ جغرافیایی وجود دارد؟

سؤال فرعی: چه استفادهٔ مدیریتی می‌توان از داده‌کاوی بر روی کاربران داده‌های دریایی در پژوهشگاه ملی اقیانوس شناسی و علوم جوی نمود؟

ارتباط بین کاربران داده و ویژگی‌های فوق برای مدیران سازمان‌های دریایی مشخص نمی‌باشد. خوشه‌بندی کاربران در سازمان‌ها توسط داده‌کاوی باعث می‌شود که مدیران بتوانند اطلاعات آنها را تحلیل نمایند، الگوی کاربران را پیدا کنند، و ارتباط بین سازمان و کاربران بهبود یابد تا در نهایت، کاربران احساس رضایت نمایند (Lai 2009).

اطلاع از این موضوع نیز از آنجا اهمیت دارد که خوشه‌بندی کاربران بر اساس تقاضای آنها و شناخت نیاز کاربران موجب خواهد شد که با مدیریت مناسب داده‌های دریایی، در جهت جلب رضایت کاربران گام برداشته شده و سرمایه‌های سازمانی به‌درستی استفاده شود. بدین‌ترتیب، از تمرکز روی جمع‌آوری داده‌های کم‌اهمیت جلوگیری می‌شود و برای داده‌های بااهمیت‌تر، سرمایه‌گذاری بیشتری خواهد شد. نبودِ این شناخت باعث خواهد شد سرمایه‌های سازمان در پی داده‌های ناکارآمد اتلاف شود.

۲. پیشینهٔ پژوهش

پیشینهٔ پژوهش در دو بخش ارائه می‌گردد: در پیشینهٔ نظری به تعاریف پایه، مباحث کشف دانش، داده‌کاوی، خوشه‌بندی و الگوریتم TwoStep و مفاهیم داده‌های دریایی پرداخته شده است و در پیشینهٔ تجربی، پژوهش‌های مرتبط انجام‌شده با استفاده از تکنیک داده‌کاوی و الگوریتم‌های خوشه‌بندی مرور شده است.

1. data type
2. data set
3. parameters
4. data source

۲-۱. پیشینهٔ نظری

۲-۱-۱. تعاریف پایه

داده[1]: به هرگونه نماد[2]، عدد، رقم، کاراکتر، رشته و یا سیگنال که معنای خاصی را به ذهن القاء نکند، داده گفته می‌شود.

اطلاعات[3]: چنانچه در کنار عدد، کاراکتر و یا هر عنصر داده‌ای، رشته‌ای به‌عنوان توصیف‌کننده وجود داشته باشد، دادهٔ ابتدایی به اطلاعات تبدیل خواهد شد.

دانش[4]: وجود یک رابطه میان دو عنصر اطلاعاتی، مبین دانشی در آن زمینه است.

خرد[5]: خرد، عالی‌ترین سطح بینش است که توسط علائم و نمادهای قراردادی تبیین می‌شود.

پایهٔ اصلی این تحقیق دو دسته داده می‌باشد: داده‌های خام جمع‌آوری‌شده از ایستگاه‌های اندازه‌گیری دریایی که پس از پردازش توسط متخصص علوم دریایی، از آنها اطلاعات حاصل می‌شود. همچنین داده‌های جمع‌آوری‌شده از پروفایل کاربران که شامل مشخصات حقیقی/ حقوقی و درخواست آنها می‌باشد.

۲-۱-۲. کشف دانش از پایگاه داده[6]

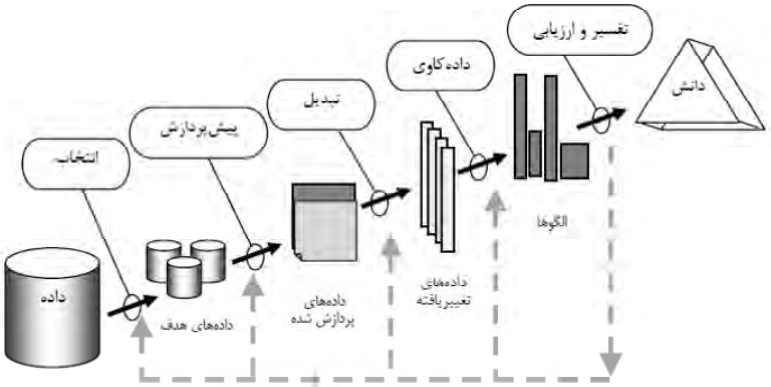
کشف دانش از پایگاه داده در واقع فرایند تشخیص الگوها[7] و مدل‌های موجود در داده‌هاست؛ الگوها و مدل‌هایی که معتبر، بدیع[8]، بالقوه مفید و کاملاً قابل فهم هستند (Fayyad, Shapiro, & Smyth 1996). به‌عبارت دیگر، هدف این فرایند یافتن الگوها و یا مدل‌های جالب موجود در پایگاه داده‌هاست که در میان حجم عظیمی از داده‌ها مخفی هستند (غضنفری ۱۳۸۷).

فرایند کشف دانش از پایگاه داده شامل مراحل کلی انتخاب، پیش‌پردازش، تبدیل، به‌کارگیری روش‌های داده‌کاوی و ارزیابی محصولات حاصل از داده‌کاوی مطابق با

---

1. Data
2. Symbol
3. Information
4. Knowledge
5. Wisdom
6. Knowledge Discovery from Database (KDD)
7. patterns
8. novel

شکل ۱ می‌باشد (Fayyad, Shapiro, & Smyth 1996).



شکل ۱. فرایند کشف دانش از پایگاه داده ۲-۱-۳. داده‌کاوی

داده‌کاوی یکی از گام‌های فرایند کشف دانش از پایگاه داده است که شامل به‌کارگیری تحلیل داده و الگوریتم‌های کشف‌شده می‌باشد و با پذیرش محدودیت‌های محاسباتی، الگوهای خاصی را تولید می‌کند (Fayyad, Shapiro, & Smyth, 1996).

۲-۱-٤. خوشه‌بندی[1]

خوشه‌بندی، به عمل تقسیم جمعیت ناهمگن به تعدادی از زیرمجموعه‌ها یا خوشه‌های همگن گفته می‌شود. در خوشه‌بندی هیچ دستهٔ از پیش تعیین‌شده‌ای وجود ندارد و داده‌ها صرفاً بر اساس تشابه گروه‌بندی می‌شوند و عناوین هر گروه نیز توسط کاربر تعیین می‌گردد (شهرابی ۱۳۹۲). خوشه‌بندی به این شکل انجام می‌شود که رکوردهایی که بیشترین شباهت را با یکدیگر دارند، در یک خوشه قرار می‌گیرند. در نتیجه، داده‌های موجود در خوشه‌های متفاوت کمترین شباهت را با یکدیگر خواهند داشت. هدف در همهٔ الگوریتم‌های خوشه‌بندی کمینه‌کردن فاصلهٔ درون‌خوشه‌ای و بیشینه‌نمودن فاصلهٔ بین خوشه‌ای می‌باشد. عملکرد خوب یک الگوریتم خوشه‌بندی زمانی

1. clustering

محرز می‌شود که تا حد امکان خوشه‌ها را از یکدیگر دور کند و علاوه بر آن، رکوردهای موجود در یک خوشه بیشترین شباهت را با یکدیگر دارا باشند. طبق تحقیقی که در سال ۲۰۱۲ انجام شده، یکی از کاربردهای خوشه‌بندی، کشف الگو در میان داده‌ها می‌باشد (Liao, Chu, & Hsiao 2012).

۲-۱-٤-۱. الگوریتم TwoStep

این الگوریتم از یک روش خوشه‌بندی دو مرحله‌ای استفاده می‌کند. مرحلهٔ اول، با یک‌بار گذر از داده‌ها، آنها را در مجموعهٔ قابل قبولی از زیرخوشه‌ها فشرده می‌کند. قدم دوم، از یک روش خوشه‌بندی سلسله‌مراتبی به‌منظور ادغام تکاملی این زیرخوشه‌ها به خوشه‌های بزرگتر بهره می‌برد (شهرابی ۱۳۹۲). یکی از مزایای این الگوریتم اجرا بر روی مجموعهٔ داده‌های بزرگ است (همان) و آنها را با کارایی زیاد اداره می‌کند (علیزاده ۱۳۹۲).

۲-۱-۵. دادهٔ دریایی

در کشورهای مختلف، داده‌ها و اطلاعات اقیانوسی از منابع مختلف، یا به‌طور مستقیم از اندازه‌گیری‌ها و مطالعات مستمر و یا موردی حاصل می‌شوند و یا ممکن است از دیگر مراکز فعال در حیطهٔ علوم اقیانوسی به‌دست آیند. بنابراین، منابع دادهٔ دریایی یا ملی و یا بین‌المللی هستند. ماهیت منابع داده‌ها و همچنین نوع داده‌ای که حاصل می‌شود، متفاوت است. عمدهٔ منابع تهیهٔ داده‌های دریایی عبارت‌اند از: اندازه‌گیری‌های ایستگاهی که ممکن است به‌صورت پروفایل‌های قائم، سری زمانی و یا نمونه‌برداری باشند. ایستگاه‌های اندازه‌گیری، محل‌هایی هستند که در آنها سنجش پارامترهای دریایی به‌صورت نقطه‌ای انجام می‌شود و داده‌هایی از نوع فیزیک دریا، شیمی دریا و ... برداشت می‌شوند (مرادی ۱۳۸۹). داده‌های دریایی در قالب مجموعهٔ داده‌هایی توسط کشورهای مختلف در محیط‌های آبی در جهان برداشت می‌شوند که معمولاً به‌نام متولی سازمان برداشت‌کنندهٔ داده یا منطقهٔ جغرافیایی یا گشت دریایی نام‌گذاری می‌شوند. این مجموعهٔ داده‌ها علاوه بر پارامترهای اندازه‌گیری مانند دما و شوری آب شامل مشخصات مکانی ایستگاه اندازه‌گیری می‌باشد که منطقهٔ جغرافیایی ایستگاه، مانند دریای خزر، خلیج فارس و دریای عمان را مشخص می‌کند.

۲-۲. پیشینهٔ تجربی

در ساختار پیشینهٔ تجربی به پژوهش‌هایی پرداخته شده که در داخل و خارج از کشور در زمینهٔ به‌کارگیری الگوریتم‌های داده‌کاوی در موضوع خوشه‌بندی کاربران انجام شده است.

در سال ۱۳۸۵، تحقیقی با عنوان «خوشه‌بندی و برچسب‌زنی کاربران وب‌سایت با استفاده از روش‌های داده‌کاوی» انجام شد. در این تحقیق خوشه‌های بازدیدکنندگان شناسایی شدند و یک تقسیم‌بندی از کاربران بر اساس رفتارشان به‌دست آمد (خاکباز ۱۳۸۵). تحقیق حاضر با هدف استفاده از این تقسیم‌بندی و به‌منظور هدفمندسازی تبلیغات و سازماندهی مجدد سایت انجام شده است و در این تحقیق الگوریتم خوشه‌بندی کوهونن[1] به‌کار گرفته شده است.

در سال ۱۳۸۷، تحقیقی برای دستیابی به ویژگی‌های تقاضای خدمت در بیمهٔ ایران با استفاده از داده‌کاوی انجام شد. در این تحقیق، قواعدی جهت تحلیل رفتار کاربران خدمات بیمهٔ ایران از داده‌های ثبت‌شده در رکوردهای کاربران استخراج گردید تا مدیران به سمت عرضهٔ خدمات مناسب با خواسته‌های کاربران حرکت نمایند. در این تحقیق از خوشه‌بندی و قانون همبستگی استفاده شده است (حسینی ۱۳۸۷).

در سال ۱۳۸۸، تحقیقی به‌منظور به‌کارگیری تکنیک داده‌کاوی برای بهبود مدیریت ارتباط با کاربران در بانک پارسیان انجام شد. با استفاده از داده‌کاوی، روابط بین سرویس‌های مورد علاقهٔ هر خوشه از کاربران کشف گردید تا در به‌کارگیری استراتژی مناسب برای ارائهٔ سرویس مناسب به کاربران استفاده شود. در این تحقیق از الگوریتم خوشه‌بندی کوهونن و الگوریتم Apriori قوانین همبستگی استفاده شده است (حسینی بامکان ۱۳۸۸).

در سال ۱۳۸۹، تحقیقی با عنوان طراحی مدل دسته‌بندی کاربران «بانک همراه» با استفاده از داده‌کاوی انجام شد. در این تحقیق با استفاده از تکنیک‌های درخت تصمیم، شبکه‌های عصبی مصنوعی و بیز ساده، مشتریان را دسته‌بندی کرده و با توجه به قوانین استخراج‌شده پیشنهادهایی جهت خدمات بهتر ارائه شده است (قنبری ۱۳۸۹).

1. Kohonen

در سال ۱۳۸۹، تحقیقی در رابطه با سرویس‌دهی بهینهٔ یک شرکت سرویس‌دهندهٔ اینترنت مبتنی بر داده‌کاوی انجام شد تا با استفاده از داده‌های تماس پروفایل کاربران، کارایی شرکت را بهبود بخشد. در این تحقیق از الگوریتم درخت تصمیم C5.0 و تلفیق آن با الگوریتم خوشه‌بندی K-Means به‌منظور تخمین مدت زمان حل مشکلات مشترکان استفاده شده است (کریمی ۱۳۸۹).

در سال ۱۳۸۹، تحقیقی به‌منظور بخش‌بندی گردشگران بین‌المللی در ایران با استفاده از تکنیک داده‌کاوی انجام شد تا با شناخت نیازهای گردشگران به مزیت رقابتی دست یابد. در این تحقیق از روش‌های مختلف خوشه‌بندی شامل الگوریتم K-Means، Kohonen و TwoStep استفاده شده است (حاجی‌بابا ۱۳۸۹).

در سال ۱۳۸۹، تحقیقی به‌منظور شناسایی کاربران بانکداری الکترونیکی بر اساس میزان استفادهٔ آنها از ابزارهای الکترونیکی انجام شد. در این تحقیق ضمن استفاده از الگوریتم K-Means و پروفایل کاربران و بررسی میزان استفادهٔ آنها، خوشه‌بندی انجام شده است (امیری، علیزاده، و لطیفی ۱۳۸۹).

در سال ۱۳۹۰، تحقیقی به‌منظور بخش‌بندی مشتریان لپ‌تاپ در ایران با استفاده از تحلیل توأم و خوشه‌بندی دومرحله‌ای انجام شد. این تحقیق قصد داشت که بر اساس نیازهای مشتریان به آنها خدمات ارائه دهد (اقدایی ۱۳۹۰).

در تحقیقی در سال ۲۰۱۰، از الگوریتم خوشه‌بندی دومرحله‌ای برای شناسایی پروفایل کاربران بانک استفاده شد تا با مدیریت مؤثر کاربران، منابع را بهتر مدیریت کنند. اشاره شده است از آنجا که داده‌های این تحقیق هم‌پیوسته[1] و هم‌رده‌ای[2] هستند، الگوریتم دومرحله‌ای بهترین الگوریتم خوشه‌بندی است (Schiopu 2010).

در سال ۱۳۹۲، تحقیقی انجام شده و با استفاده از الگوریتم K-Means دانشجویان رشتهٔ کامپیوتر را بر اساس میزان علاقه به درس برنامه‌نویسی خوشه‌بندی کرده است (شریفی، حسینی، و گلیج ۱۳۹۲).

در تحقیقی در سال ۱۳۹۳، با استفاده از الگوریتم دومرحله‌ای به خوشه‌بندی کاربران تلفن همراه پرداخته شد. این تحقیق متذکر شده است که در روش سنتی خوشه‌بندی

1. continuous
2. categorical

معمولاً از روش‌های غیرسلسله‌مراتبی مانند K-Means استفاده می‌شده که برای داده‌های بازه‌ای مناسب هستند. لیکن، الگوریتم دومرحله‌ای برای داده‌های رده‌ای مناسب می‌باشند و همچنین نیازی به تعیین اولیهٔ تعداد خوشه‌ها ندارد (نبی‌زاده و نجمی ۱۳۹۳).

نمونه‌هایی از مطالعات انجام‌شده در زمینهٔ داده‌کاوی بر روی کاربران و الگوریتم‌های آن در جدول ۱ آورده شده است.

جدول ۱. مطالب پیشینهٔ تحقیق

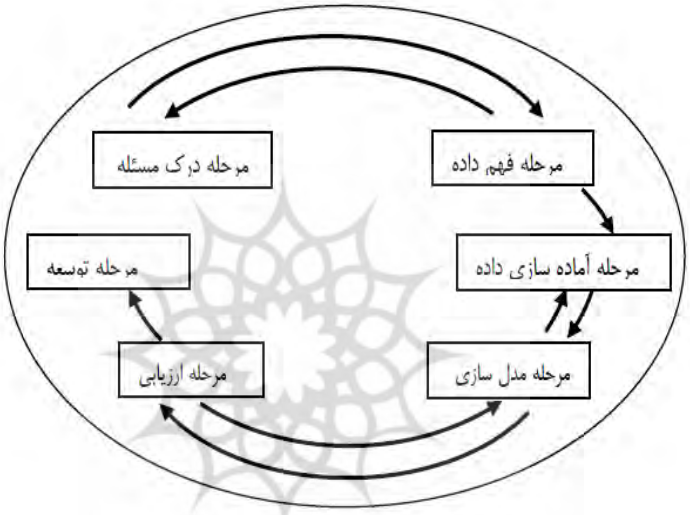
نام محقق	سال	موضوع	تکنیک داده‌کاوی	الگوریتم
خاکباز	۱۳۸۵	داده‌کاوی کاربران وب‌سایت	خوشه‌بندی	کوهونن
حسینی	۱۳۸۷	تحلیل رفتار کاربران خدمات بیمه	خوشه‌بندی	
حسینی بامکان	۱۳۸۸	یافتن روابط بین سرویس‌های مورد علاقهٔ کاربران بانک	خوشه‌بندی قوانین همبستگی	کوهونن Apriori
قنبری	۱۳۸۹	دسته‌بندی کاربران بانک همراه	درخت تصمیم، شبکه‌های عصبی مصنوعی و بیز ساده	
کریمی	۱۳۸۹	سرویس‌دهی بهینه یک شرکت سرویس‌دهنده اینترنت	درخت تصمیم خوشه‌بندی	C5.0 K-Means
حاجی بابا	۱۳۸۹	شناخت نیازهای گردشگران بین‌المللی	خوشه‌بندی	K-Means Kohonen TwoStep

نام محقق	سال	موضوع	تکنیک داده‌کاوی	الگوریتم
امیری و علیزاده و لطیفی	۱۳۸۹	شناسایی کاربران بانکداری الکترونیکی بر اساس میزان استفادهٔ آنها از ابزارهای الکترونیکی	خوشه‌بندی	K-Means
اقدایی	۱۳۹۰	یافتن نیازهای مشتریان لپ‌تاپ بر اساس نیازهایشان	خوشه‌بندی	TwoStep
Schiopu	۲۰۱۰	شناسایی پروفایل کاربران بانک	خوشه‌بندی	TwoStep
شریفی و حسینی و گلیج	۱۳۹۲	یافتن میزان علاقه به درس برنامه‌نویسی دانشجویان رشتهٔ کامپیوتر	خوشه‌بندی	K-Means
نبی‌زاده و نجمی	۱۳۹۳	خوشه‌بندی کاربران تلفن همراه	خوشه‌بندی	TwoStep
پژوهش حاضر	۱۳۹۳	داده‌کاوی در کاربران داده‌های دریایی	خوشه‌بندی	TwoStep

با مرور پژوهش‌های انجام‌شده می‌توان دریافت که از تکنیک داده‌کاوی در خوشه‌بندی کاربران دریایی تحقیقی انجام نشده است، در حالی که همان‌گونه که در مقدمه بیان شد، اطلاع از این موضوع حائز اهمیت است. در حوزه‌های دیگر مطالعات محدودی انجام شده است که تکنیک غالب در این تحقیقات، خوشه‌بندی و الگوریتم‌های مورد استفادهٔ آنها K-Means، Kohonen و TwoStep بوده است. جنبهٔ نوآوری این تحقیق «خوشه‌بندی کاربران داده‌های دریایی با استفاده از تکنیک داده‌کاوی برای اولین بار در ایران» می‌باشد.

۳. روش پژوهش

در این تحقیق از مدل استاندارد CRISP-DM[1] که یکی از روش‌های بسیار قوی در داده‌کاوی است و توسط شرکت‌های دایملر کرایسلر[2]، اس‌پی‌اس‌اس[3] و ان‌سی‌آر[4] توسعه‌یافته، استفاده شده است. این متدولوژی از گام‌های شناخت سیستم، درک داده، آماده‌سازی داده‌ها، مدل‌سازی، ارزیابی و توسعهٔ سیستم مطابق شکل ۲ تشکیل شده است.



شکل ۲. مراحل متدولوژی CRISP-DM

برای داشتن تحقیقی موفق در زمینهٔ داده‌کاوی، باید متخصص داده‌کاوی از توان و تجربهٔ متخصص کسب‌وکار در تمام فرایند داده‌کاوی بهره‌مند باشد (رادفر، نظافتی، و یوسفی اصل ۱۳۹۳).

1. Cross Industry Standard Process for Data Mining
2. Daimler Crysler
3. SPSS
4. NCR

۳-۱. شناخت کسب‌وکار (سیستم):

با توجه به رشد روزافزون تولید داده‌های دریایی و به‌کارگیری آنها توسط متخصصان علوم دریایی، شناخت رابطهٔ میان کاربر با داده‌ها، می‌تواند هم برای کاربران و هم برای مدیران و تصمیم‌گیران مفید باشد. کاربران با توجه به دسته‌ای که در آن قرار می‌گیرند، می‌توانند از توصیه‌های مدیران دریایی برای استفاده از داده‌های دریایی بهره‌مند شوند و انتخاب صحیحی از داده‌ها برای انجام کارهای تحقیقاتی خود داشته باشند.

هدف داده‌کاوی در این تحقیق «خوشه‌بندی» و نوع آن «توصیفی» است؛ بدین معنا که اطلاعات جدید و غیربدیهی را بر اساس مجموعه‌ای از داده‌های موجود ارائه می‌دهد و هدف کلی آن درک و شناخت سیستم‌های تجزیه و تحلیل‌شده با استفاده از الگوها و روابط موجود است (رادفر، نظافتی، و یوسفی اصل ۱۳۹۳).

کاربران با توجه به خوشه‌های استخراج‌شده می‌توانند از توصیه‌های مدیران دریایی برای استفاده از داده‌های دریایی بهره‌مند شوند و انتخاب صحیحی از داده‌ها برای انجام کارهای تحقیقاتی خود داشته باشند. مدیران نیز با شناخت تقاضای کاربران می‌توانند برنامهٔ استراتژیک خود را پایه‌ریزی نمایند.

۳-۲. درک داده

در این تحقیق از میان جامعهٔ آماری ۶۰۰ کاربر دریایی، اطلاعات و پروفایل ۵۰۰ کاربر که از سال ۱۳۸۶ تا ۱۳۹۳ در پژوهشگاه ملی اقیانوس‌شناسی و علوم جوی جمع‌آوری‌شده، استفاده شده است. این نمونهٔ آماری به‌دلیل تشابه درخواست‌ها انتخاب شده‌اند و کاربرانی که تقاضاهای خاصی به لحاظ منبع داده، نوع داده، مجموعهٔ داده، پارامتر و منطقهٔ جغرافیایی داشتند، در نظر گرفته نشده‌اند. کاربران داده‌های دریایی را می‌توان به‌طور کلی به سه دستهٔ عمده تقسیم کرد: دانشجویان کارشناسی ارشد و دکترای رشته‌های علوم دریایی، پژوهشگران و سازمان‌های درگیر در فعالیت‌های دریایی.

در این تحقیق برای ارتباط کاربران با داده‌های درخواستی خود، از میان جامعهٔ آماری ۴۰ مجموعهٔ داده، ۱۸ مجموعهٔ داده که محیط‌های آبی ایران در شمال و جنوب را پوشش می‌دادند، از منابع ملی و بین‌المللی با نوع داده‌های فیزیک دریا، شیمی دریا، آلودگی دریا و هواشناسی با پارامترهای مرتبط در نظر گرفته شده است. مجموعهٔ

داده‌هایی که خارج از محدودهٔ آبی کشور ایران بودند و همچنین، مجموعهٔ داده‌هایی که محدودهٔ کوچکی را پوشش می‌دادند، در نظر گرفته نشده‌اند.

۳-۳. آماده‌سازی داده

در این مرحله انتخاب داده، با انتخاب جداول، رکوردها و ویژگی‌ها انجام شده است. در مرحلهٔ پاکسازی داده‌ها مشکلات داده‌های پرت[1] و مقادیر مفقوده[2] برطرف شده‌اند. در مرحلهٔ یکپارچه‌سازی *داده‌ها* کلیهٔ داده‌ها از جداول مختلف در یک پایگاه داده مجتمع شده‌اند و در کاهش داده فیلدهای نوع کاربر، منبع داده، نوع داده، مجموعهٔ داده، پارامتر و منطقهٔ جغرافیایی از میان فیلدهای مجموعهٔ داده‌ها استخراج شده‌اند.

۳-٤. مدل‌سازی

در این مرحله از تکنیک خوشه‌بندی برای مدل‌سازی استفاده شده است. به‌منظور مدل‌سازی از نرم‌افزار IBM SPSS Modeler که یکی از نرم‌افزارهای مشهور در زمینهٔ داده‌کاوی است، استفاده شده است. با استفاده از این نرم‌افزار سه نوع الگوریتم K-Means، Kohonen و TwoStep برای مدل‌سازی به‌کار گرفته شده است. هر یک از این سه الگوریتم بر روی داده‌ها اعمال شد و با توجه به شاخص‌های ارزیابی بهترین الگوریتم انتخاب شد.

۳-۵. ارزیابی

ارزیابی کیفیت خوشه‌بندی، میزان برتری یک خوشه‌بندی نسبت به خوشه‌بندی‌های دیگر به‌وسیلهٔ الگوریتم‌های متفاوت خوشه‌بندی یا الگوریتم‌های مشابه، ولی با مقدار پارامترهای متفاوت می‌باشد (سفیداری، کدخدائی، و شریفی ۱۳۹۲). در این تحقیق از شاخص سیلوئت[3] به همراه زمان اجرا و تعداد خوشه‌ها استفاده شده است.

شاخص سیلوئت:

شاخص تراکم و جدایی سیلوئت با مقادیر ضعیف[4]، متوسط[5] و خوب[6] نشان داده

1. outlier data
2. missing value
3. Silhouette
4. poor
5. fair
6. good

می‌شود. میانگین مقدار شاخص سیلوئت برای ارزیابی اعتبار خوشه‌بندی و همچنین برای تصمیم‌گیری در مورد انتخاب تعداد کلاس‌های بهینه مورد استفاده قرار می‌گیرد که این میزان بر اساس دوری و نزدیکی مشاهدات و خوشه‌ها به یکدیگر محاسبه می‌شود. مقدار S(i) با استفاده از فرمول ۱ قابل محاسبه است:

$$S(i)=\frac{(b(i)-a(i))}{\max(a(i),b(i))}$$ فرمول ۱

a(i) میانگین فاصله بین مشاهده i با سایر مشاهدات در یک خوشهٔ مشابه و b(i) میانگین فاصلهٔ مشاهده i با تمام مشاهدات در خوشه‌های دیگر می‌باشد. بر اساس فرمول بالا مقدار S(i) بین −۱ و +۱ قرار دارد. اگر S(i) به +۱ نزدیک‌تر باشد، به این معنی است که خوشه‌بندی نمونهٔ خوب صورت گرفته است و خوشهٔ پیشنهادشده برای نمونهٔ مورد نظر مناسب می‌باشد، ولی اگر S(i) به −۱ نزدیک‌تر باشد، به این معنی است که خوشه‌بندی نمونه، به‌خوبی انجام نشده و خوشهٔ پیشنهادشده برای دادهٔ مورد نظر نامناسب می‌باشد.

با مقایسهٔ سه معیار شاخص سیلوئت، زمان اجرا و تعداد خوشه‌ها، الگوریتم TwoStep مطابق جدول ۲ برای خوشه‌بندی نوع کاربران با داده‌های درخواستی انتخاب شد.

جدول ۲. معیارهای ارزیابی در الگوریتم TwoStep برای خوشه‌بندی نوع کاربر و منبع داده

معیار / مدل‌سازی	زمان اجرا (ثانیه)	تعداد خوشه‌ها	شاخص سیلوئت
نوع کاربر با منبع داده	۱	۴	۰/۹
نوع کاربر با نوع داده	۱	۵	۱
نوع کاربر با مجموعهٔ داده	۲	۵	۰/۶
نوع کاربر با پارامتر	۵	۱۰	۰/۳
نوع کاربر با منطقهٔ جغرافیایی	۱	۵	۰/۸

یکی از مزایای این الگوریتم، اجرا بر روی مجموعهٔ داده‌های بزرگ است (شهرابی ۱۳۹۲)، و آنها را با کارایی زیاد اداره می‌کند (علیزاده ۱۳۹۲). از دیگر نقاط قوت آن این است که قادر به مدیریت داده‌هایی با انواع مختلف فیلدهاست.

در تحقیقی که در سال ۱۳۹۰ انجام شده، از الگوریتم TwoStep برای خوشه‌بندی خلیج چابهار بر اساس چگالی استفاده شده است (وفایی و همکاران ۱۳۹۲).

٤. تجزیه و تحلیل یافته‌ها

در این بخش از پژوهش، الگوریتم پذیرفته‌شده در بخش مدل‌سازی تحقیق روی داده‌ها اعمال شده و نتایج حاصل از آنها در قالب نمودار و جدول ارائه شده است. الگوریتم TwoStep نوعی تحلیل خوشه‌بندی است که از فیلد هدف استفاده نمی‌کند. این خوشه‌بندی یک روش خوشه‌بندی دو مرحله‌ای است. گام اول، با یک گذر از روی داده‌های ورودی، آنها را به زیرخوشه‌های قابل مدیریت فشرده می‌کند. گام دوم، از یک روش خوشه‌بندی سلسله‌مراتبی به‌منظور ادغام تکاملی این زیرخوشه‌ها به خوشه‌های بزرگ‌تر بهره می‌برد.

◇ **نتایج خوشه‌بندی نوع کاربر با منبع داده**

با انتخاب الگوریتم TwoStep به‌عنوان الگوریتم بهینه در زمان ارزیابی مدل، نتایج خوشه‌بندی مطابق جدول ۳ به‌دست آمد.

جدول ۳. نتایج خوشه‌بندی نوع کاربر با منبع داده با الگوریتم TwoStep

نام خوشه به‌ترتیب اندازه	نوع کاربر	منبع داده	سایز خوشه‌ها به درصد
خوشهٔ ۱	دانشجو	بین‌المللی	۵۲/۲
خوشهٔ ۳	سازمان	بین‌المللی	۲۱/۱
خوشهٔ ۴	پژوهشگر	بین‌المللی	۱۶/۹
خوشهٔ ۲	دانشجو	ملی	۹/۸

چهار خوشه بر اساس نوع کاربر با منبع داده به‌دست آمد که بر اساس میزان تقاضا به‌ترتیب از زیاد به کم به شرح زیر بودند:

خوشهٔ ۱: نوع کاربر دانشجو با منبع دادهٔ بین‌المللی

خوشهٔ ۳: نوع کاربر سازمان با منبع دادهٔ بین‌المللی

خوشهٔ ٤: نوع کاربر پژوهشگر با منبع دادهٔ بین‌المللی

خوشهٔ ۲: نوع کاربر دانشجو با منبع دادهٔ ملی

از تحلیل خوشه‌بندی‌های فوق، نتایج و الگوهای زیر به‌دست آمد:

نتیجهٔ ۱: منابع بین‌المللی چون در اندازه‌گیری‌های خود، مناطق وسیع‌تر و ابزارهای

اندازه‌گیری دقیق‌تری را دربرمی‌گیرند، تقاضای بیشتری در میان سه نوع کاربر به نسبت سایر منابع داده برخوردار هستند.

نتیجهٔ ۲: منابع ملی و ملی- بین‌المللی از تقاضای کمتری برخوردار هستند.

◇ **نتایج خوشه‌بندی نوع کاربر با نوع داده**

با انتخاب الگوریتم TwoStep به‌عنوان الگوریتم بهینه در زمان ارزیابی مدل، نتایج خوشه‌بندی مطابق جدول ۴ به‌دست آمد.

جدول ٤. نتایج خوشه‌بندی نوع کاربر با نوع داده با الگوریتم TwoStep

نام خوشه به‌ترتیب اندازه	نوع کاربر	نوع داده	سایز خوشه‌ها به درصد
خوشهٔ ۲	دانشجو	شیمی دریا	۳۵/۳
خوشهٔ ۱	دانشجو	فیزیک دریا	۲۰/۹
خوشهٔ ۳	پژوهشگر	شیمی دریا	۱۵/۷
خوشهٔ ۵	سازمان	شیمی دریا	۱۵/۱
خوشهٔ ۴	سازمان	فیزیک دریا	۱۳/۱

پنج خوشه بر اساس نوع کاربر با نوع داده به‌دست آمد که بر اساس میزان تقاضا به‌ترتیب از زیاد به کم به شرح زیر بودند:

خوشهٔ ۲: نوع کاربر دانشجو با نوع دادهٔ شیمی دریا

خوشهٔ ۱: نوع کاربر دانشجو با نوع دادهٔ فیزیک دریا

خوشهٔ ۳: نوع کاربر پژوهشگر با نوع دادهٔ شیمی دریا

خوشهٔ ٥: نوع کاربر سازمان با نوع دادهٔ شیمی دریا

خوشهٔ ٤: نوع کاربر سازمان با نوع دادهٔ فیزیک دریا

از تحلیل خوشه‌بندی‌های فوق، نتایج و الگوهای زیر به‌دست آمد:

نتیجهٔ ۱: بیشترین تقاضای داده از نظر نوع کاربر با دانشجویان با نوع دادهٔ شیمی دریا می‌باشد.

نتیجهٔ ۲: بیشترین تقاضای داده از نظر هر ۳ نوع کاربر داده دانشجویان، پژوهشگران و سازمان‌ها با نوع دادهٔ شیمی دریا می‌باشد.

نتیجهٔ ۳: از آنجا که مجموعهٔ داده‌های شیمی دریا و فیزیک دریا به‌عنوان نوع داده‌های اساسی در مطالعات دریایی هستند، از اهمیت بیشتری برای دانشجویان، پژوهشگران و سازمان‌ها برخوردار می‌باشند.

نتیجهٔ ٤: اندازه‌گیری‌های نوع دادهٔ زمین شناسی، ژئوفیزیک و ... پیچیده‌تر و مشکل‌تر از نوع دادهٔ شیمی و فیزیک می‌باشد.

◇ **نتایج خوشه‌بندی نوع کاربر با مجموعهٔ داده**

با انتخاب الگوریتم TwoStep به‌عنوان الگوریتم بهینه در زمان ارزیابی مدل، نتایج خوشه‌بندی مطابق جدول ۵ به‌دست آمد.

جدول ۵. نتایج خوشه‌بندی نوع کاربر با مجموعهٔ داده با الگوریتم TwoStep

نام خوشه به‌ترتیب اندازه	نوع کاربر	مجموعه داده	سایز خوشه‌ها به درصد
خوشهٔ ۲	دانشجو	پایگاه دادهٔ اقیانوسی جهانی	۲۵/۱
خوشهٔ ۱	سازمان	پایگاه دادهٔ اقیانوسی جهانی	۲۴/۷
خوشهٔ ۳	پژوهشگر	پایگاه دادهٔ اقیانوسی جهانی	۱۸/۱
خوشهٔ ۵	دانشجو	پایگاه دادهٔ مرکز دادهٔ هند	۱۷/۲
خوشهٔ ۴	دانشجو	پایگاه دادهٔ اقیانوسی جهانی	۱۴/۹

پنج خوشه بر اساس نوع کاربر با مجموعهٔ داده به‌دست آمد که بر اساس میزان تقاضا به‌ترتیب از زیاد به کم به شرح زیر بودند:

خوشهٔ ۲: نوع کاربر دانشجو با مجموعهٔ دادهٔ پایگاه دادهٔ اقیانوسی جهانی

خوشهٔ ۱: نوع کاربر سازمان با مجموعه دادهٔ پایگاه دادهٔ اقیانوسی جهانی

خوشهٔ ۳: نوع کاربر پژوهشگر با مجموعه دادهٔ پایگاه دادهٔ اقیانوسی جهانی

خوشهٔ ۵: نوع کاربر دانشجو با مجموعه دادهٔ پایگاه دادهٔ مرکز دادهٔ هند

خوشهٔ ٤: نوع کاربر دانشجو با مجموعه دادهٔ پایگاه دادهٔ اقیانوسی جهانی

از تحلیل خوشه‌بندی‌های فوق، نتایج و الگوهای زیر به‌دست آمد:

نتیجهٔ ۱: بیشترین تقاضای داده از نظر نوع کاربر با دانشجویان با مجموعهٔ داده «پایگاه دادهٔ

اقیانوسی جهانی» در دریای خزر می‌باشد.

نتیجهٔ ۲: بیشترین تقاضای داده از نظر هر ۳ نوع کاربر داده دانشجویان، سازمان‌ها و پژوهشگران با مجموعهٔ داده «پایگاه دادهٔ اقیانوسی جهانی» می‌باشد.

نتیجهٔ ۳: «پایگاه دادهٔ اقیانوسی جهانی» به‌دلیل وسعت مناطق جغرافیایی، تاریخی‌بودن و تعدد پارامترهای اندازه‌گیری، مجموعهٔ دادهٔ غنی برای کاربران مختلف می‌باشد.

◇ **نتایج خوشه‌بندی نوع کاربر با پارامتر**

با انتخاب الگوریتم TwoStep به‌عنوان الگوریتم بهینه در زمان ارزیابی مدل، نتایج خوشه‌بندی مطابق جدول ۶ به‌دست آمد.

جدول ٦. نتایج خوشه‌بندی نوع کاربر با پارامتر با الگوریتم TwoStep

نام خوشه به‌ترتیب اندازه	نوع کاربر	پارامتر	سایز خوشه‌ها به درصد
خوشهٔ ۶	سازمان	نیترات	۱۴/۰
خوشهٔ ۲	دانشجو	دمای آب	۱۳/۸
خوشهٔ ۱۰	پژوهشگر	شوری	۱۰/۸
خوشهٔ ۱	دانشجو	شوری	۹/۷
خوشهٔ ۳	دانشجو	اکسیژن	۹/۵
خوشهٔ ۹	سازمان	دمای آب	۸/۹
خوشهٔ ۷	دانشجو	آلکالینیتی	۸/۶
خوشهٔ ۵	دانشجو	نیترات	۸/۵
خوشهٔ ۴	دانشجو	سیلیکات	۸/۳
خوشهٔ ۸	پژوهشگر	دمای آب	۷/۹

ده خوشه بر اساس نوع کاربر با پارامتر به‌دست آمد که بر اساس میزان تقاضا به‌ترتیب از زیاد به کم به شرح زیر بودند:

خوشهٔ ٦: نوع کاربر سازمان با پارامتر نیترات

خوشهٔ ۲: نوع کاربر دانشجو با پارامتر دمای آب

خوشهٔ ۱۰: نوع کاربر پژوهشگر با پارامتر شوری

خوشهٔ ۱: نوع کاربر دانشجو با پارامتر شوری

خوشهٔ ۳: نوع کاربر دانشجو با پارامتر اکسیژن

خوشهٔ ۹: نوع کاربر سازمان با پارامتر دمای آب

خوشهٔ ۷: نوع کاربر دانشجو با پارامتر آلکالینیتی

خوشهٔ ۵: نوع کاربر دانشجو با پارامتر نیترات

خوشهٔ ٤: نوع کاربر دانشجو با پارامتر سیلیکات

خوشهٔ ۸: نوع کاربر پژوهشگر با پارامتر دمای آب

نتیجهٔ ۱: بیشترین تقاضای داده از نظر نوع کاربر سازمان‌ها با پارامتر نیترات و سپس نوع کاربر دانشجو با پارامتر دمای آب و شوری می‌باشد.

نتیجهٔ ۲: پارامترهای دمای آب و شوری از پارامترهای اصلی مطالعات دریایی می‌باشند.

نتیجهٔ ۳: دسترسی و اندازه‌گیری سایر پارامترها به سختی انجام می‌شود.

نتیجهٔ ٤: پارامترهای شیمی دریا و فیزیک دریا از اهمیت بیشتری برای کاربران داده برخوردار می‌باشد.

◇ **نتایج خوشه‌بندی نوع کاربر با منطقهٔ جغرافیایی**

با انتخاب الگوریتم TwoStep به‌عنوان الگوریتم بهینه در زمان ارزیابی مدل، نتایج خوشه‌بندی مطابق جدول ۷ به‌دست آمد.

جدول ۷. نتایج خوشه‌بندی نوع کاربر با منطقهٔ جغرافیایی با الگوریتم TwoStep

نام خوشه به‌ترتیب اندازه	نوع کاربر	منطقهٔ جغرافیایی	سایز خوشه‌ها به درصد
خوشهٔ ۳	دانشجو	خلیج فارس	۱۴/۰
خوشهٔ ۱	سازمان	خلیج فارس	۱۳/۸
خوشهٔ ۴	دانشجو	دریای خزر	۱۰/۸
خوشهٔ ۲	پژوهشگر	دریای خزر	۹/۷
خوشهٔ ۵	دانشجو	دریای عمان	۹/۵

پنج خوشه بر اساس نوع کاربر با منطقهٔ جغرافیایی به‌دست آمد که بر اساس میزان تقاضا به‌ترتیب از زیاد به کم به شرح زیر بودند:

خوشهٔ ۳: نوع کاربر دانشجو با منطقهٔ خلیج فارس

خوشهٔ ۱: نوع کاربر سازمان با منطقهٔ خلیج فارس

خوشهٔ ٤: نوع کاربر دانشجو با منطقهٔ دریای خزر

خوشهٔ ۲: نوع کاربر پژوهشگر با منطقهٔ دریای خزر

خوشهٔ ٥: نوع کاربر دانشجو با منطقهٔ دریای عمان

نتیجهٔ ۱: بیشترین تقاضای داده از نظر نوع کاربر دانشجویان و سازمان‌ها با منطقهٔ خلیج فارس می‌باشد.

نتیجهٔ ۲: خلیج فارس به‌دلیل موقعیت اقتصادی، استراتژیک و ژئوپلیتیک از اهمیت بیشتری میان کاربران داده برخوردار است.

نتیجهٔ ۳: تقاضای داده از نظر ۲ نوع کاربر دانشجویان و پژوهشگران با منطقهٔ دریای خزر در اولویت بعدی می‌باشد.

نتیجهٔ ٤: دسترسی به داده‌های دریای خزر از طریق کشورهای حاشیهٔ آن مشکل‌تر می‌باشد.

نتیجهٔ ٥: دسترسی به خلیج فارس و دریای خزر از نظر کار تحقیقاتی نسبت به دریای عمان ساده‌تر است.

نتیجهٔ ٦: دسترسی به داده‌های دریای عمان به‌دلیل عمق زیاد و عدم همکاری کشورهای پاکستان و هند در تبادل داده مشکل می‌باشد.

نتیجهٔ ۷: در دریای عمان، شکاف کار تحقیقاتی وجود دارد.

◇ **بررسی اعتبار الگوهای کشف‌شده**

داده‌های تحقیق با استفاده از روش داده‌کاوی به الگوهای جدید تبدیل می‌شود. در واقع، الگوهای جدید همان اطلاعات هستند. با ارائهٔ این اطلاعات به خبرگان علوم دریایی و تأیید نتایج توسط آنها این اطلاعات تبدیل به دانش می‌شود. اعتبار نتایج به‌دست آمده، توسط ۵ نفر از اعضای هیئت علمی پژوهشگاه تأیید شد.

۵. نتیجه‌گیری

از تحلیل خوشه‌بندی نوع کاربران با داده‌های درخواستی آنها، با کمک فرد خبره نتایج و الگوهایی در علوم دریایی به شرح زیر به‌دست آمد:

منابع بین‌المللی چون در اندازه‌گیری‌های خود، مناطق وسیع‌تر و ابزارهای اندازه‌گیری دقیق‌تری را دربرمی‌گیرند، از تقاضای بیشتری در میان سه نوع کاربر به نسبت منابع ملی برخوردار هستند. بیشترین تقاضای داده از نظر هر سه نوع کاربر داده دانشجویان، پژوهشگران و سازمان‌ها با نوع دادهٔ شیمی دریا می‌باشد. از آنجا که مجموعهٔ داده‌های شیمی دریا و فیزیک دریا به‌عنوان نوع داده‌های اساسی در مطالعات دریایی هستند، از اهمیت بیشتری برای کاربران برخوردار می‌باشند. بیشترین تقاضای داده از نظر هر سه نوع کاربر داده با مجموعهٔ دادهٔ «پایگاه دادهٔ اقیانوسی جهانی» می‌باشد. این پایگاه داده، به‌دلیل وسعت مناطق جغرافیایی، تاریخی‌بودن داده‌ها و تعدد پارامترهای اندازه‌گیری، مجموعهٔ دادهٔ غنی برای کاربران مختلف می‌باشد. بیشترین تقاضای داده از نظر نوع کاربر سازمان‌ها با پارامتر نیترات و سپس نوع کاربر دانشجو با پارامتر دمای آب و شوری می‌باشد. پارامترهای دمای آب و شوری از پارامترهای اصلی مطالعات دریایی می‌باشند. دسترسی و اندازه‌گیری سایر پارامترها، به‌سختی انجام می‌شود. بیشترین تقاضای داده از نظر نوع کاربر دانشجویان و سازمان‌ها با منطقهٔ خلیج فارس می‌باشد. خلیج فارس به‌دلیل موقعیت اقتصادی، استراتژیک و ژئوپلیتیک از اهمیت بیشتری میان کاربران داده برخوردار است. دسترسی به داده‌های دریای خزر از طریق کشورهای حاشیهٔ آن مشکل‌تر می‌باشد. دسترسی به داده‌های دریای عمان به‌دلیل عمق زیاد و عدم همکاری کشورهای پاکستان و هند در تبادل داده مشکل می‌باشد. در دریای عمان، شکاف کار تحقیقاتی وجود دارد.

فواید این الگوها برای پژوهشگاه ملی اقیانوس‌شناسی و علوم جوی این است که می‌توانند سرمایه‌گذاری مالی و انسانی بیشتری برای جمع‌آوری داده‌های بین‌المللی، شیمی دریا، فیزیک دریا، مجموعهٔ دادهٔ «پایگاه دادهٔ اقیانوسی جهانی»، پارامترهای دما، شوری و منطقهٔ خلیج فارس نمایند و از سرمایه‌گذاری برای سایر منابع، نوع داده، مجموعهٔ داده‌ها، پارامترها و مناطق دریایی پرهیز نمایند.

در این قسمت با توجه به تجزیه و تحلیل داده‌ها و نتایج حاصل از طرح به سؤالات تحقیق پاسخ داده می‌شود.

سؤال اصلی: چه الگویی بین کاربران داده با نوع داده، مجموعهٔ داده، پارامترهای داده، منبع داده و منطقهٔ جغرافیایی وجود دارد؟

پاسخ سؤال اصلی: ازخوشه‌های به‌دست آمده و نتایج حاصل از آنها می‌توان دریافت که به‌طور خلاصه، الگوی بین *کاربران با منبع داده* این است که منابع بین‌المللی، تقاضای بیشتری در میان سه نوع کاربر به نسبت سایر منابع داده برخوردار هستند. الگوی بین کاربران با نوع داده این است که بیشترین تقاضای داده از نظر هر ۳ نوع کاربر داده با نوع دادهٔ شیمی دریا می‌باشد. الگوی بین کاربران با مجموعهٔ داده این است که بیشترین تقاضای داده از نظر هر ۳ نوع کاربر داده با مجموعهٔ دادهٔ «پایگاه دادهٔ اقیانوسی جهانی» می‌باشد. الگوی بین کاربران با پارامتر این است که پارامترهای شیمی دریا و فیزیک دریا از اهمیت بیشتری برای کاربران داده، برخوردار می‌باشند. و در نهایت، الگوی بین کاربران با مناطق جغرافیایی این است که خلیج فارس به‌دلیل موقعیت اقتصادی، استراتژیک و ژئوپلیتیک از اهمیت بیشتری میان کاربران داده برخوردار است.

سؤال فرعی: چه استفادهٔ مدیریتی می‌توان از داده‌کاوی بر روی کاربران دادهٔ دریایی در پژوهشگاه ملی اقیانوس‌شناسی و علوم جوی نمود؟

پاسخ سؤال فرعی: با توجه به الگوهای به‌دست آمده، مدیران ارشد که درگیر فعالیت‌های تصمیم‌گیری هستند و همچنین، مدیران فناوری اطلاعات که در فرایند مدیریت داده‌های دریایی فعالیت می‌کنند، قادر خواهند بود به‌درستی در مورد داده‌های موجود خود و برنامه‌ریزی برای جمع‌آوری داده در آینده تصمیم‌گیری کنند و درک بهتری از نیازهای کاربران خود داشته باشند. اطلاع از این موضوع نیز از آنجا اهمیت دارد که شناخت نیاز کاربران، موجب خواهد شد با مدیریت مناسب داده‌های دریایی، در جهت جلب رضایت کاربران گام برداشته شود و سرمایه‌های سازمانی به‌درستی استفاده شود. بدین‌ترتیب، از تمرکز روی جمع‌آوری داده‌های کم‌استفاده یا بدون استفاده جلوگیری خواهد شد.

با توجه به نتایج فوق، به مدیران ارشد پژوهشگاه ملی اقیانوس‌شناسی و علوم جوی و سایر ارگان‌های دریایی کشور که درگیر فعالیت‌های تصمیم‌گیری هستند، همچنین مدیران فناوری اطلاعات که در فرایند مدیریت داده‌های دریایی فعالیت می‌کنند، و کاربران داده که استفاده‌کنندهٔ داده‌ها هستند، پیشنهاد می‌شود:

۱. علاوه بر تقویت و بروزرسانی بانک اطلاعات دریایی پژوهشگاه با داده‌های منابع بین‌المللی، به کارهای میدانی، برگزاری گشت‌های دریایی داخلی، و تقویت منابع دادهٔ ملی توسط ارگان‌های دریایی پرداخته شود.
۲. به‌دلیل تقاضای بیشتر نوع دادهٔ شیمی دریا، فیزیک دریا و پارامترهای نیترات، دمای آب، و شوری به اندازه‌گیری مستمر این داده‌ها پرداخته شود. همچنین، به جمع‌آوری داده در مناطق خلیج فارس و دریای خزر و تقویت بانک اطلاعاتی خود در این زمینه بپردازند و هزینهٔ کمتری در جمع‌آوری داده در مناطق دریای عمان و دریای عرب صرف نمایند.
۳. دانشجویان تحصیلات تکمیلی، پژوهشگران و سازمان‌ها برای دسترسی آسان‌تر و بیشتر به منابع تحقیقاتی می‌توانند نوع دادهٔ مورد مطالعهٔ خود را شیمی و فیزیک دریا و پارامترهای مورد مطالعهٔ خود را نیترات، دمای آب، و شوری و مناطق مورد مطالعهٔ خود را در خلیج فارس و دریای خزر تعریف نمایند.
۴. دانشجویان تحصیلات تکمیلی، پژوهشگران و سازمان‌ها به‌منظور انجام کار تحقیقاتی بدیع و تازه می‌توانند نوع دادهٔ مورد مطالعهٔ خود را بیولوژی دریا، آلودگی دریا، زمین‌شناسی و ژئوفیزیک دریایی و پارامتر مورد مطالعه خود را از میان این نوع داده‌ها و مناطق مورد مطالعهٔ خود را در دریای عمان و دریای عرب تعریف نمایند.
۵. پژوهشگاه به دریافت مستمر مجموعهٔ دادهٔ «پایگاه دادهٔ اقیانوسی جهانی» که از اهمیت به‌سزایی در میان کاربران داده برخوردار است و به‌روزرسانی بانک اطلاعاتی با مجموعهٔ دادهٔ فوق اقدام نماید.
۶. پیشنهاد می‌شود، سازمان‌های دریایی به ایجاد مجموعهٔ دادهٔ تاریخی ملی، مشابه «پایگاه دادهٔ اقیانوسی جهانی» (که منبع آن بین‌المللی می‌باشد) بپردازند.

در انتها از آنجا که داده‌های این تحقیق متعلق به داده‌های پژوهشگاه ملی اقیانوس‌شناسی و علوم جوی است، پیشنهاد می‌شود تحقیقی با روش‌های به‌کاررفته در این تحقیق، در سایر سازمان‌های دریایی انجام گیرد و نتایج آن با نتایج این تحقیق مقایسه شود. در صورتی که مدیران ارشد پژوهشگاه ملی اقیانوس‌شناسی و علوم جوی به مطالعهٔ الگوهای تقاضای کاربران داده تمایل داشته باشند، می‌توان درخواست‌های کاربران را در

سال‌های پس از ۱۳۹۳ نیز بررسی کرده و از نتایج آنها در تدوین چشم‌انداز و برنامه‌های راهبردی پژوهشگاه استفاده نمایند.

فهرست منابع

اقدایی، محمدحسن. ۱۳۹۰. *بخش‌بندی منافعی مشتریان لپ‌تاپ در ایران با استفاده از تحلیل توأم و خوشه‌بندی دو مرحله‌ای*. آمل. دانشگاه شمال.

امیری، فهیمه، سمیه علیزاده، و فریبا لطیفی. ۱۳۸۹. *تحلیل و شناسایی رفتار مشتریان بانکداری الکترونیکی با استفاده از داده‌کاوی*. چهارمین کنفرانس بین‌المللی مدیریت بازاریابی، تهران.

حاجی بابا، هما. ۱۳۸۹. *بخش‌بندی گردشگران بین‌المللی در ایران با استفاده از تکنیک داده‌کاوی*. تهران: دانشگاه الزهرا.

حسینی، زهرا. ۱۳۸۷. *به‌کارگیری فرایند داده‌کاوی برای دستیابی به ویژگی‌های تقاضای خدمت*. مشهد: دانشگاه فردوسی مشهد.

حسینی بامکان، سیدمجتبی. ۱۳۸۸. *به‌کارگیری تکنیک‌های داده‌کاوی جهت بهبود مدیریت ارتباط با مشتری در صنعت بانکداری*. تهران: دانشگاه علامه طباطبایی.

خاکباز، محمدحسین. ۱۳۸۵. *خوشه‌بندی و برچسب‌زنی کاربران وب‌سایت با استفاده از روش‌های داده‌کاوی*. تهران: دانشگاه تربیت‌مدرس.

رادفر، رضا، نوید نظافتی، و سعید یوسفی اصلی. ۱۳۹۳. طبقه‌بندی مشتریان بانک با کمک الگوریتم‌های داده‌کاوی. *مجله مدیریت فناوری اطلاعات*. دوره ۶. شماره ۱. صص ۷۱-۹۰.

سفیداری، ابراهیم، علی کدخدائی، و محمد شریفی. ۱۳۹۲. مقایسهٔ روش‌های شبکهٔ عصبی خودسازنده و آنالیز خوشه‌ای برای ارزیابی مقدار کربن آلی در سازندهای محتوی هیدروکربن با استفاده از سیستم‌های هوشمند. *پژوهش نفت*. سال بیست‌وسوم. شماره ۷۵. صفحه ۱۱۷-۱۳۰

شریفی، ابوصالح محمد، سیدمهدی حسینی، و ابراهیم گلیج. ۱۳۹۲. *استفاده از الگوریتم K-Means برای خوشه‌بندی دانشجویان رشتهٔ کامپیوتر بر اساس میزان علاقه‌مندی و عوامل مؤثر بر یادگیری دروس برنامه‌نویسی*. اولین کنفرانس ملی نوآوری در مهندسی کامپیوتر و فناوری اطلاعات. تنکابن.

شهرابی، جمال ۱۳۹۲. *داده‌کاوی*. تهران: *جهاد دانشگاهی*. واحد صنعتی امیرکبیر.

علیزاده، سمیه ۱۳۹۲. *داده‌کاوی و کشف دانش گام به گام با نرم‌افزار*. تهران: دانشگاه صنعتی خواجه نصیرالدین طوسی.

غضنفری، مهدی ۱۳۸۷. *داده‌کاوی و کشف دانش*. تهران: دانشگاه علم و صنعت.

قنبری، محمدحسام. ۱۳۸۹. *طراحی مدل دسته‌بندی کاربران بانک همراه با استفاده از الگوریتم داده‌کاوی*. تهران: دانشگاه تربیت‌مدرس.

کریمی، فائزه. ۱۳۸۹. *ارائهٔ راهکار مناسب در راستای سرویس‌دهی بهینهٔ مرکز تماس یک شرکت ISP ایرانی مبتنی بر داده‌کاوی*. تهران: دانشگاه تربیت‌مدرس.

مرادی، مسعود. ۱۳۸۹. *مدیریت داده‌های اقیانوس‌شناسی*. تهران: مسعود مرادی.

نبی‌زاده، محمد، و منوچهر نجمی. ۱۳۹۲. *ارائهٔ روشی نوین در خوشه‌بندی بازار خدمات تلفن همراه با استفاده از داده‌کاوی*. تهران: دانشگاه صنعتی شریف.

وفایی، بشرا، وحید چگینی، عباس سقایی، و مجتبی عظام. ۱۳۹۲. دسته‌بندی توده‌های آب در خلیج چابهار با استفاده از روش خوشه‌بندی. *اقیانوس شناسی*. سال چهارم. شماره ۱۳. صفحه ۱۱-۱۹

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. From data mining to knowledge discovery in databases. *AI magazine* 17 (3): 37 .

Lai, X.-a. 2009. *Segmentation Study on Enterprise Customers Based on Data Mining Technology*. Paper presented at the First International Workshop on Database Technology and Applications.

Liao, S.-H., P-H. Chu, and P-Y. Hsiao. 2012. Data mining techniques and applications–A decade review from 2000 to 2011. *Expert Systems with Applications* 39 (12): 11303-11311 .

Schiopu, D. 2010. Applying TwoStep cluster analysis for identifying bank customers' profile. *Stiinte Economice* 1.LXII, 66-75.