ISSN: 2745-3673
http://www.Rcsj.ir

مجله‌ی علمی پژوهش در علوم رایانه (سال چهارم)
شماره ۱۵ / پاییز ۱۳۹۸ / ص ۳۵- ۲۳

آماده سازی داده برای فرایندکاوی

فریدون شمس علیئی[۱]، لیلا حیدری[۲]، محمود نشاطی[۳]

[۱] دانشیار دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، ایران.

[۲] کارشناسی ارشد مهندسی فناوری اطلاعات، دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، ایران.

[۳] استادیار دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، ایران.

نام نویسنده مسئول:
فریدون شمس علیئی

چکیده

با ظهور کامپیوتر و بکارگیری آن در امور جاری کسب و کارها، حجم بزرگی از داده‌ها ایجاد شد. این داده‌ها که با اهداف مختلفی جمع‌آوری می‌شدند، به مرور مورد توجه دانشمندان قرار گرفتند. آنها به فکر افتادند تا با پردازش این انباره‌های عظیم داده، روابط پنهان میانشان را کشف نموده و از دانش حاصله در آینده کسب و کارها استفاده نمایند. امروزه انواع تکنیک‌های پردازش داده وجود دارد. یکی ازآنها فرایندکاوی است. اما داده‌های موجود در انباره‌های داده خام هستند و برای پردازش نیاز به آماده‌سازی دارند. مقالات مختلفی در این خصوص موجود است که هر یک روالی را برای آماده‌سازی داده به منظوری خاص، ارائه نموده است. این مقاله پس از بررسی مراحل اصلی آماده‌سازی داده که لازم است در تمام پردازش‌های مبتنی بر داده انجام شود؛ چرخه آماده‌سازی داده برای فرایندکاوی را با توجه به چالش‌های موجود در این زمینه، پیشنهاد می‌کند. سپس چرخه پیشنهادی، با انجام یک مطالعه موردی بر روی داده‌های موجود از فرایند کنترل کیفیت درخواست‌های کارت هوشمند ملی ایران، ارزیابی شده است.

واژگان کلیدی: آماده سازی داده، پیش پردازش داده، فرایندکاوی، آماده سازی داده برای فرایندکاوی.

مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۱۵، پاییز ۱۳۹۸، ص ۳۵- ۲۳

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

پس از پیدایش کامپیوتر و بکارگیری آن در امور جاری سازمان‌ها و مراکز کسب و کار، حجم بزرگی از داده‌ها در حافظه‌های این کامپیوترها نگهداری شد. بنا به قانون مور[1]، حجم داده‌ها، ظرفیت دیسک‌ها، قدرت محاسباتی کامپیوترها و ... هر ساله دو برابر شده است. داده‌های نگهداری شده در بانک‌های اطلاعاتی، با اهداف مختلفی جمع‌آوری شده‌اند. کشف روابط و ساختارهای پنهان میان داده‌ها، ما را به عصاره این انبوه داشته‌ها و دانشی گرانبها می‌رساند. براین اساس انواع تکنیک‌های کاوش مانند متن‌کاوی[2]، داده‌کاوی[3]، وب‌کاوی[4]، دانش‌کاوی[5] و فرایندکاوی[6] ایجاد شد. اما هر یک از این تکنیک‌ها نیاز به داده‌های خاصی دارند که باید از میان انبوه داده‌های موجود استخراج شوند. بنابراین جهت تحلیل داده‌ها، نیاز به شناخت و آماده‌سازی آنها، متناسب با هدف مورد نظر است.

مساله‌ای که اکنون با آن مواجه هستیم کمبود داده نیست بلکه استخراج داده صحیح، قابل اطمینان و موثر، از میان حجم انبوه داده‌های موجود در بانک‌های اطلاعاتی است. این موضوع برای افرادی که با علم داده آشنا هستند و یا با داده سروکار دارند، کاملاً آشکار است. کیفیت داده‌های استخراج شده ارتباط مستقیم با کیفیت نتایج حاصل از پردازش آنها دارد. در این خصوص وان‌درالست و همکارانش در بیانیه فرایندکاوی می‌گویند: "داده‌ها شهروندان درجه یک هستند"[۱۰]. وظیفه اصلی آماده‌سازی داده، سازماندهی داده‌ها متناسب باشکل استاندارد پردازش مورد نظر (مانند داده‌کاوی، فرایندکاوی یا ...) است.

مقالات زیادی موجود است که مراحل آماده‌سازی یا پیش پردازش داده را قبل از هر نوع پردازش مبتنی بر داده، ارائه نموده‌اند. در این مقاله چرخه‌ای از ترتیب مراحل آماده‌سازی داده برای فرایندکاوی ارائه شده است که به چالش‌های موجود در این زمینه توجه دارد. بدین منظور ابتدا علم فرایندکاوی معرفی شده است. سپس مراحل اصلی آماده سازی داده که لازم است پیش از هر پردازش مبتنی بر داده انجام شود، ارائه شده است. پس از آن به چالش‌های فرایندکاوی در زمینه آماده‌سازی داده پرداخته شده است. در نهایت چرخه آماده‌سازی داده برای فرایندکاوی ارائه شده و به کمک داده‌های مربوط به فرایند کنترل کیفیت درخواست‌های کارت هوشمند ملی ایران، مورد ارزیابی قرار گرفته است.

فرایندکاوی چیست؟

فرایندکاوی یک رویکرد مدیریت فرایند است که شامل تکنیک‌ها، ابزارها و روش‌هایی برای کشف، تحلیل و بهبود فرایندهای کسب و کار، با استفاده از داده‌های ثبت شده در نگاره‌های رویداد[7] می‌باشد. این داده‌ها، در حین اجرای فرایندهای کسب و کار، توسط سیستم‌های اطلاعاتی تولید و ثبت می‌شوند. علم فرایندکاوی از این داده‌ها برای کشف مدل‌های فرایندی که در دنیای واقعی اتفاق افتاده است، استفاده می‌کند. سپس این مدل‌ها تحلیل و بهینه‌سازی می‌شود. از آنجایی که این اطلاعات سیستمی ثبت می‌شود (نه دستی)، از درجه اطمینان بالاتری برخوردار هستند. زیرا در ثبت دستی اطلاعات احتمال خطا یا اعمال سلیقه فردی بالا می‌رود. در حقیقت با کمک تکنیک‌های فرایندکاوی می‌توان مسیری که فرایند در دنیای واقعی طی نموده است، را یافت.

فرایندکاوی حوزه تحقیقاتی نسبتاً جدید است که اواخر دهه ۱۹۹۰ توسط Cook و Wolf در زمینه تحلیل فرایندهای مهندسی نرم‌افزار ارائه شد و همچنان در حال تکامل است[۲۲]. اولین تحقیق فرایندکاوی در سال۲۰۰۳ انجام شد پس از آن تکنیک‌های فرایندکاوی روز به روز رشد کرده و ابزارهای مختلفی برای پشتیبانی از آن ایجاد شده است[۲۱]. در دهه اخیر، با فراهم شدن دسترسی به داده‌های موجود در نگاره رویداد، روش‌های فرایندکاوی به بلوغ رسیدند. درحال حاضر گروه تحقیقاتی وان‌درالست و همکارانش در دانشگاه ایندهون هلند تحقیقات گسترده‌ای را درخصوص فرایندکاوی آغاز نموده‌اند. ابزار ProM که یک ابزار قوی برای کاوش فرایند است، از مهمترین دست آوردهای این گروه است.

علم فرایندکاوی، برمبنای دو علم مدیریت فرایندها و تحلیل داده‌ها بنا شده است. این علم شکاف بین روش‌های تجزیه و تحلیل مدل فرایندی، بدون توجه به داده‌ها (مانند شبیه سازی)، و تکنیک‌های تجزیه و تحلیل کلاسیک داده‌ها (مانند داده‌کاوی)، را پر کرده است[۲۰]. کاوش فرایند می‌تواند یکی از مهم‌ترین ابزارها برای مجریان کسب و کارهایی باشد که نیاز به مدیریت مناسب فرایندهای عملیاتی خویش دارند. زیرا:

---

[1] Moore's law
[2] Text Mining
[3] Data Mining
[4] Web Mining
[5] Knowledge Mining
[6] Process Mining
[7] Event log

ISSN: 2745-3673
http://www.Rcsj.ir

۱. رویدادهایی که روزانه در پایگاه‌های داده ثبت می‌شود، به صورت نمایی در حال رشد هستند. این رویدادها جزئیات تاریخچه‌ی فرایندهای اجرا شده کسب و کار را دارا می‌باشند.
۲. با توجه به رقابتی شدن فضای کسب و کار و محیط در حال تغییر، از جمله تغییرات تکنولوژی، افراد و سیاست‌های کاری، نیاز به بهبود فرایندهای حرفه گریز ناپذیراست.

بنابراین مناسب است تکنیک‌های فرایندکاوی را برای مدیریت فرایندهای کسب و کار به درستی پیاده‌سازی کرد تا فرایندها به صورت مستمر ارزیابی، پایش و بهبود داده شوند.

علم فرایندکاوی شامل سه تکنیک اصلی است که عبارتند از: کشف[۸]، انطباق[۹]، ارتقاء[۱۰].

در این رویکرد، با بررسی نگاره رویداد نمونه‌هایی از رفتارهای سیستم مشاهده می‌شود. سپس فرایند مورد بررسی را مشخص نموده و فعالیت‌های آن به همراه زمان و منبع انجام دهنده فعالیت، استخراج می‌شود. حال با استفاده از تکنیک‌های کشف فرایند، قوانین و نقش‌ها را یافته و مدل‌های مربوط به جریان کار[۱۱] (روال اجرای فرایند از ابتدا تا انتها)، شبکه اجتماعی[۱۲] (نقش‌ها و تاثیرشان بر روال اجرای فرایند)، زمان و منابع اجرا کننده فرایند را استخراج می‌کند. پس از آن لازم است با تکنیک‌های انطباق، میزان مطابقت مدل کشف شده با مدل اصلی فرایند کسب و کار را بدست آورد. به این ترتیب نقایص، خطاها، انحرافات و گلوگاه‌ها کشف شده و مسیر مطلوب کاربران معین می‌گردد. در نهایت سعی می‌شود با استفاده از تکنیک‌های بهبود و ارتقاء، مدل‌های فرایندی بهبود داده شود.

مراحل اصلی آماده‌سازی داده

آماده‌سازی داده مهمترین و وقت‌گیرترین کار برای هر نوع پردازش مبتنی بر داده است. زیرا داده‌های ثبت شده در پایگاه داده، به صورت خام نگهداری می‌شود. برای تبدیل داده‌های خام به اطلاعات مورد نیاز، لازم است اعمالی روی این داده‌ها انجام شود. آماده‌سازی داده به سه دلیل اهمیت دارد:

۱. استخراج داده‌های صحیح
۲. افزایش کارایی سیستم‌های کاوش، با داده با کیفیت
۳. افزایش کیفیت نتیجه پردازش[۱۲].

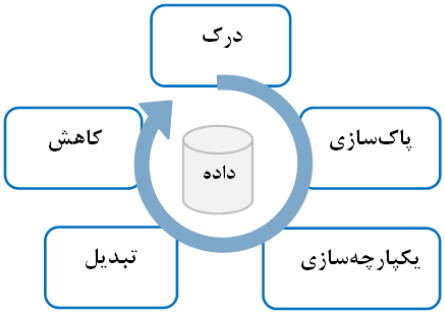
پر واضح است که فقدان داده با کیفیت به معنای فقدان کیفیت نتایج است. به عبارتی دیگر هرگز از یک ورودی نامناسب، خروجی مناسبی حاصل نمی‌شود. در مقاله "اهمیت پردازش داده‌ها"، تحقیقی در خصوص اهمیت آماده‌سازی داده‌ها نسبت به سایر گام‌های کشف دانش با روش داده‌کاوی انجام شده است. این تحقیق نشان می‌دهد که ۷۵ درصد زمان پروژه‌های داده‌کاوی صرف آماده‌سازی داده‌ها می‌شود[۴]. این موضوع در مقالات متعددی بررسی شده است که نشان می‌دهد؛ حدود ۷۰ تا ۸۰ درصد زمان پروژه‌های مبتنی بر داده، صرف آماده‌سازی یا پیش‌پردازش داده‌ها شده است. این مطلب گویای اهمیت بالای آماده‌سازی داده با کیفیت است.

در مقالات مختلف، مراحل متفاوتی برای آماده‌سازی داده، ارائه شده است. اما در تمامی آنها ۵ مرحله اصلی که در شکل (۱) مشاهده می‌شود، وجود دارد.

- درک داده[۱۳]
- پاک‌سازی داده[۱۴]
- یکپارچه‌سازی داده[۱۵]
- تبدیل داده[۱۶]
- کاهش داده[۱۷][۱][۲][۳][۵][۶][۷][۸][۱۲][۱۸].

[8]discovery
[9]conformance
[10]enhancement
[11]Work flow
[12]Social network
[13]data understanding
[14]data cleaning
[15]data integration
[16]data transformation
[17]data reduction

ISSN: 2745-3673
http://www.Rcsj.ir



شکل (۱): مراحل آماده‌سازی داده قبل از پردازش

• **درک داده**: در این مرحله داده‌های مرتبط با پروژه، شناسایی و جمع‌آوری می‌شوند. اگر داده‌ها در یک پایگاه داده متمرکز ذخیره و نگهداری شوند، کار نسبتاً آسان است. اما اگر داده‌های مورد نیاز پروژه، در پایگاه داده‌های گوناگون یا به صورت دستی روی کاغذ ثبت شده باشند؛ یافتن و جمع‌آوری آنها مستلزم دقت بالا و صرف زمان است. این مرحله بسیار مهم است. زیرا نقص در داده‌ها به معنای نقص در نتایج پردازش آنها است.

• **پاک‌سازی داده:** این مرحله به منظور دستیابی به داده‌های صحیح و سازگار انجام می‌شود تا نمایش همگونی از داده‌ها ایجاد شود. بدین منظور تلاش می‌شود مشکلات مربوط به داده‌های گم شده، نویزها، داده‌های پرت و ناسازگاری‌های بین داده‌ها، شناسایی شده و راه حلی برای رفع این مشکلات پیدا نمود. اهمیت این بخش آنقدر بالاست که هرچقدر پاک‌سازی داده‌ها بهتر انجام شود، نتایج کاوش بهتر خواهد بود. بسته به تعداد منابع، درجه ناهمگونی و کثیفی داده‌ها، ممکن است نیاز به چندین مرحله پاک‌سازی داده وجود داشته باشد[۲۵].

• **یکپارچه‌سازی داده**: عملیات یکپارچه‌سازی، به دلیل گوناگونی منابع و همپوشانی داده‌ها در منابع گوناگون، تنوع در نحوه ذخیره‌سازی داده‌ها و قابلیت پردازش تقاضاهای متفاوت در منابع مختلف انجام می‌شود[۲۶]. در این گام مشکلات مربوط به تضاد و افزونگی داده‌ها بررسی و رفع می‌گردد. بنابراین در صورتی که نگاره‌ها در منابع اطلاعاتی مختلفی ثبت شده باشند؛ نیاز به یکپارچه‌سازی دارند. بدین منظور لازم است زبان نگاره‌ها یکسان گردد. یعنی نام و نوع ویژگی‌های هر نمونه و شناسه‌ها (کدها)ی موجود در نگاره‌ها یکسان شوند. به عبارتی دیگر اگر نگاره‌ها در بانک‌های اطلاعاتی مختلفی ثبت و نگهداری شده باشند؛ احتمال دارد ویژگی‌های نگهداری شده از هر نمونه و یا شناسه‌های استفاده شده در هر نگاره با دیگری متفاوت باشد. ضروری است که این ویژگی‌ها و شناسه‌ها، پس از تشخیص، تحلیل و یکپارچه شوند. با این عمل تضادها و عدم تطابق بین داده‌ها، شناسایی و یکپارچه‌سازی می‌شود. به علاوه مشکل افزونگی داده (داده‌هایی که به صورت تکراری در منابع گوناگونی نگهداری شده‌اند) نیز رفع می‌گردد. اگر مستنداتی برای این ویژگی‌ها و شناسه‌ها وجود نداشته باشد، درک و یکپارچه‌سازی آنها کاری بسیار دشوار و زمان‌بری خواهد بود.

• **تبدیل داده**: عملیاتی همچون نرمال‌سازی، تغییر و تجمیع داده‌ها در این گام انجام می‌شود. زیرا ویژگی‌های نگهداری شده در منابع اطلاعاتی، داده‌هایی خام هستند. یعنی ویژگی‌ها، متناسب با حوزه کاری خاص و یا نتیجه کار سامانه‌ای مشخص، طراحی و نگهداری شده‌اند. این داده‌ها مناسب پردازش نیستند و لازم است به استانداری متناسب با پروژه مورد نظر تبدیل شوند. جهت نرمال‌سازی داده‌ها نیاز به تغییر ویژگی‌های موجود و یا ایجاد ویژگی‌های جدید است. روش‌های گوناگونی برای نرمال‌سازی داده وجود دارد. از جمله:

- Min-Max Normalization
- Z-score Normalization
- Decimal-scaling Normalization

در انتهای این مرحله کلیه اطلاعات مورد نیاز برای یک پروژه پردازش داده، تمیز، یکپارچه و نرمال شده و در یک بانک اطلاعاتی مشخص تجمیع می‌شود.

• **کاهش داده**: هدف از کاهش داده، دست‌یابی به حجم کوچک‌تری از داده‌ها است. یکی از مهم‌ترین دلایل کاهش داده، حجم بالای داده‌هاست که تحلیل آنها را پیچیده، زمان‌بر و گاهی غیر ممکن می‌کند. دلیل دوم، محدودیت ابزارهای پردازش داده است. کاهش داده، کاری بسیار حساس و مهم است. هدف تکنیک‌های کاهش داده، دست‌یابی به زیر مجموعه کوچکتری از انبوه داده‌هاست که خصوصیات داده اصلی را داشته باشد. دو تکنیک "انتخاب ویژگی" و "نمونه‌برداری" ازپرکاربردترین تکنیک‌های کاهش داده هستند.

ISSN: 2745-3673
http://www.Rcsj.ir

به منظور آماده‌سازی و پردازش داده برای کاربردهای مختلف، ابزارهای گوناگونی توسعه داده شده است. مانند Trifacta، SOAPnuke، RapidMiner، DataRobot، Paxata، Azure، TIBCO و ... .

مراحل آماده‌سازی داده برای فرایندکاوی

آماده‌سازی داده برای فریندکاوی که یکی از تکنیک‌های پردازش داده است، علاوه بر چالش‌هایی که تمام پردازش‌های مبتنی بر داده دارند، با چالش‌هایی دیگری نیز روبرو است که نیاز به تحلیل و بررسی دقیق دارد. از جمله چالش‌های آماده‌سازی داده برای فرایندکاوی عبارتند از:

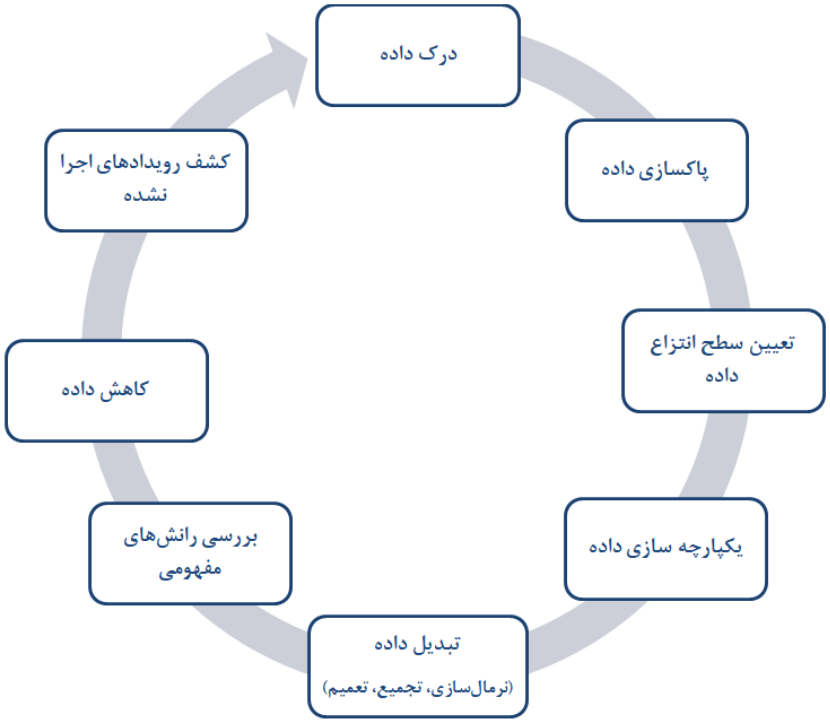
- "شی محور" بودن داده‌های موجود در نگاره‌ها (داده‌های ثبت شده در نگاره‌ها "فرایند محور" نیستند).
- یافتن، ادغام کردن و تمیز دادن داده‌های رویداد با توجه به منابع اطلاعاتی توزیع شده، تنوع در شناسه‌گذاری[۱۸]، ناکامل بودن نگاره رویداد، داده‌های اضافه و داده‌های پرت
- سایز نگاره (نگاره‌ها بزرگ و پیچیده / نگاره‌ها کوچک با داده‌های ناکافی)
- یافتن رویدادها مربوط به هر نمونه، شباهت بین نمونه‌ها، رویدادهای منحصر به فرد و مسیرهای منحصر به فرد
- یافتن رویدادهایی از فرایند، که هرگز اجرا نشده‌اند
- متفاوت بودن سطح انتزاع داده‌های ذخیره شده در نگاره‌ها
- رانش مفهومی[۱۹](تغییر فرایند در حال اجرا) که از مهم‌ترین چالش‌های کاوش فرایند از نگاره‌ها است[۱۰].

بنابراین اولین چالش در آماده‌سازی داده برای فرایندکاوی، پس از یافتن نگاره‌ها، جستجوی ویژگی‌های مورد نیاز برای کاوش فرایند است. بعد از آن ضروریست که شناخت نمونه‌ها و رویدادهای مربوط به آنها، جهت یافتن مسیرهایی که هر نمونه طی کرده است، انجام شود تا نگاره‌های شی محور را تبدیل به نگاره‌های فرایند محور نمود. تعیین سطح انتزاع نگاره‌ها و کشف رویدادهایی که هرگز اجرا نشده‌اند، از دیگر اعمالی است که انجام آن در مرحله آماده‌سازی داده، اجتناب ناپذیر است. زیرا سطح انتزاع داده‌ها، در تعیین سطح انتزاع مدل فرایندی که قرار است برمبنای این داده‌ها تهیه شود موثر است. علاوه بر این اگر رویدادی هرگز اجرا نشده باشد، در مدل فرایندی کشف شده از نگاره نیز نمایش داده نمی‌شود. مهم‌تر از موارد فوق یافتن رانش‌های مفهومی است. زیرا در صورت تغییر روند اجرای فرایند در حال اجرا، مدل فرایندی کشف شده از نگاره، ترکیبی از دو یا چند روند اجرا، خواهد بود. اهمیت این موضوع به حدی زیاد است که می‌تواند نمودار گردش‌کار فرایند را تحت تاثیر قرار داده و پروژه فرایندکاوی را از هدفش دور نماید.

این مقاله با بررسی چالش‌های فوق، ترتیب مراحل آماده سازی داده برای فرایندکاوی را پیشنهاد می‌دهد. شکل(۲) چرخه‌ی پیشنهادی این مقاله را نشان می‌دهد.

[18]Coding
[19]Concept Drift

ISSN: 2745-3673
http://www.Rcsj.ir



شکل (۲): چرخه آماده سازی داده برای فرایندکاوی

این چرخه شامل ۸ مرحله است که عبارتند از: درک داده، پاک‌سازی داده، تعیین سطح انتزاع داده، یکپارچه‌سازی داده، تبدیل داده، بررسی رانش‌های مفهومی، کاهش داده و کشف رویدادهای اجرا نشده. در ادامه به تشریح هر یک از این مراحل پرداخته شده است.

درک داده: در این گام نگاره‌های رویداد شناسایی شده و داده‌های موجود از فرایند، تحلیل و بررسی می‌شود. سپس ویژگی‌های مورد نیاز برای کاوش فرایند استخراج می‌گردد. این ویژگی‌ها عبارتند از: نمونه (پرونده یا مورد)، رویداد، فعالیت، زمان انجام فعالیت و منبع انجام دهنده فعالیت[۱۹]. در برخی از فرایندها هزینه انجام فعالیت‌ها نیز نگهداری می‌شود که در بهبود فرایند می‌تواند مفید باشد. همچنین ممکن است زمان شروع و پایان یک فعالیت به صورت جداگانه نگهداری شود. در این صورت محاسبه زمان دقیق اجرای هر فعالیت امکان پذیر است. اما رویدادها لزوماً در یک فایل نگاره اختصاصی ذخیره نشده‌اند. ممکن است رویدادها در جداول متفاوت پایگاه داده، نگاره‌های تراکنش، نگاره‌های پیغام، آرشیو پستی و سایر منابع داده ذخیره شده باشند. به علاوه در فرایندهای بین سازمانی، هر بخش از نگاره در یک سازمان ثبت و نگهداری می‌شود. بنابراین شناسایی این نگاره‌ها و جمع‌آوری آنها گامی پرچالش و زمان‌بر در آماده‌سازی داده برای فرایندکاوی است. مهمتر از قالب ذخیره‌سازی، کیفیت داده‌های رویداد است[۱۰]. لازم است داده‌هایی که جمع‌آوری می‌شوند، کامل، قابل اعتماد، خوش تعریف و امن باشند[۱۰].

پاک‌سازی داده: در این گام همانند دیگر پردازش‌های مبتنی بر داده، جهت دستیابی به داده‌های صحیح و سازگار و نمایشی همگون از داده‌ها، پاک‌سازی داده انجام می‌شود. بدین ترتیب داده‌های تکراری حذف شده و مشکلات مربوط به داده‌های گم شده، نویزها، داده‌های پرت و ناسازگاری‌های بین داده‌ها، رفع می‌شود.

تعیین سطح انتزاع داده: در برخی از نگاره‌ها سطح انتزاع داده‌ها بسیار پایین است. این مطلب بدین معنی است که تعداد رویدادهای ثبت شده از یک فرایند، بیش از حد زیاد است. به عنوان مثال: رویدادهایی مانند ورود/خروج کاربر به سامانه، باز/بسته شدن یک صفحه خاص از سامانه توسط کاربر و ... نیز ثبت می‌شود. در این صورت علاوه بر بزرگ شدن حجم نگاره، سطح انتزاع داده‌ها نیز پایین می‌آید. خیلی از این داده‌ها برای پروژه فرایندکاوی کاربرد ندارد. بر عکس در برخی نگاره‌ها سطح انتزاع داده‌ها به حدی بالاست که ممکن است بعضی از رویدادها ثبت نشده باشد. تعیین سطح انتزاع مناسب باعث تمرکز بر روی رویدادهای اصلی فرایند شده و مدل فرایندی خواناتری را ایجاد می‌کند. سطح انتزاع باید طوری انتخاب شود که رویدادهای مهم فرایند حذف نشوند. به منظور تعیین سطح انتزاع داده‌های موجود در نگاره(های) یک فرایند، لازم است به ماموریت‌ها و اهداف سازمان یا شرکت مجری، توجه نمود.

یکپارچه‌سازی داده:در صورتی که نگاره‌های مربوط به یک فرایند در منابع اطلاعاتی مختلفی ثبت و نگهداری شده باشد، نیاز به یکپارچه‌سازی وجود دارد تا مشکلات مربوط به تضاد و افزونگی داده‌ها مرتفع شود. به عبارتی دیگر، در یکپارچه‌سازی داده‌ها زبان نگاره‌ها

ISSN: 2745-3673
http://www.Rcsj.ir

یکسان می‌شود. بدین منظور لازم است نام و نوع ویژگی‌های مربوط به نمونه‌ها و شناسه‌های بکار گرفته شده در نگاره‌های مختلف، پس از تشخیص، تحلیل و یکپارچه شوند. اگر مستندات کافی برای معرفی نگاره‌ها موجود نباشد؛ درک و یکپارچه‌سازی آنها کاری بسیار دشوار و زمان‌بری است.

تبدیل داده‌ها: در این گام داده‌های مورد نیاز پروژه فرایندکاوی، از منابع اطلاعاتی گوناگون تجمیع شده و نرمال‌سازی می‌شوند. به عبارتی دیگر داده‌های خام موجود در نگاره‌ها برای دستیابی به مدل دقیقی از فرایند اجرا شده، مناسب نیستند. لازم است این داده‌ها نرمال‌سازی شده و ویژگی‌های مورد نیاز استخراج شوند. در این راستا گاهی لازم است ویژگی‌های موجود تغییر کند. به عنوان مثال چند ویژگی با هم ترکیب شده و ویژگی جدیدی بسازند. در انتهای این گام، کلیه اطلاعات مورد نیاز، پس از تمیز، یکپارچه و نرمال‌سازی شدن، در یک بانک اطلاعاتی مشخص تجمیع می‌شود.

بررسی رانش‌های مفهومی: رانش مفهومی زمانی اتفاق می‌افتد که روند اجرای یک فرایند در حال اجرا، ناگهان تغییر می‌کند. این تغییر به دلایل مختلفی اتفاق می‌افتد. تغییر نیازمندی‌ها، تغییر افراد، تغییر تکنولوژی، تغییر در خدمت یا محصول، تغییر محیط داخلی یا خارجی سازمان، ضرورت ساده‌سازی و ... می‌توانند از علل تغییر روند اجرای فرایند باشند. بنابراین لازم است روند اجرای فرایند بررسی شود تا در صورت وجود رانش(های) مفهومی، آنها را شناسایی نمود. اگر مستنداتی در این خصوص وجود داشته باشد، کار کاوش در نگاره، ساده‌تر است. می‌توان کاوش را در هر یک از بازه‌های زمانی رانش‌های مفهومی انجام داد. در غیر این صورت، ممکن است لازم باشد مدل فرایندی را در بازه‌های مختلف زمان اجرای آن، تهیه و باهم مقایسه نمود. در صورت عدم تشخیص زمان رانش‌های مفهومی یک فرایند، مدل کاوش شده از نگاره همه مسیرهای اجرا شده را در برخواهد داشت. برای رفع این مشکل می‌توان کاوش فرایند از نگاره را در بازه‌های زمانی کوتاه و نزدیک به حال، با اطمینان به عدم وقوع رانش مفهومی انجام داد. مثلا یک ماه یا یک سال اخیر که روند اجرای فرایند هیچ تغییری نداشته است.

کاهش داده: کوچک کردن حجم نگاره، به دلیل کاهش پیچیدگی داده‌ها و محدودیت ابزارهای فرایندکاوی انجام می‌شود. معمولا داده‌های موجود در نگاره‌ها مختص به یک فرایند نیستند. بنابراین لازم است ابتدا داده‌های مربوط به فرایند مورد نظر شناسایی و تفکیک شود. سپس نوع نمونه‌ها با استفاده از پرسش و پاسخ، تجزیه و تحلیل شود تا چرخه حیات نمونه‌ها مشخص شود[۱۰]. حال اگر تعداد نمونه‌ها بیش از محدودیت ابزار مورد استفاده باشد؛ لازم است یکی از تکنیک‌های کاهش داده بکار گرفته شود. دو تکنیک پرکاربرد در کاهش داده، حذف ویژگی و نمونه‌گیری است. با توجه به اینکه تکنیک‌های فرایندکاوی، ویژگی‌های مشخصی را نیاز دارند، حذف ویژگی امکانپذیر نیست. اما نمونه برداری تحت شرایطی خاص می‌تواند اتفاق افتد. از آنجایی که برای کاوش فرایند، چرخه حیات نمونه‌ها اهمیت دارد، دو حالت برای نمونه برداری امکان‌پذیر است.

- حالت اول، می‌توان نگاره را براساس یک ویژگی دسته بندی نموده و کاوش فرایند را برای یک دسته خاص انجام داد و یا هر دسته را جداگانه کاوش نمود. به عنوان مثال یک فرایند سازمانی که در گستره وسیع جغرافیایی اجرا می‌شود را می‌توان براساس تقسیمات جغرافیایی دسته بندی کرده و هر دسته را جداگانه تحلیل نمود.
- حالت دوم: نگاره را براساس نمونه‌های تکمیل شده در بازه‌های زمانی مختلف دسته بندی نموده و رفتار فرایند را در یک بازه زمانی مشخص تحلیل کرد.

کشف رویدادهای اجرا نشده: پس از کاهش داده لازم است رویدادهایی که اجرا نشده‌اند و در نگاره نمونه وجود ندارند را پیدا کرد. زیرا رویدادی که در نگاره وجود نداشته باشد، در مدل کشف شده از نگاره نیز نمایش داده نمی‌شود. بدین منظور با مقایسه لیست رویدادهای موجود در مدل اصلی فرایند (متناسب با نمونه برداری انجام شده) و رویدادهای ثبت شده در نگاره، می‌توان رویدادهایی که اصلا اجرا نشده‌اند را یافت. این موضوع در زمان تطبیق مدل فرایندی کشف شده از نگاره با مدل اصلی فرایند، حائز اهمیت است.

در انتهای این چرخه، نگاره‌ای تولید شده است که کلیه اطلاعات مورد نیاز برای کشف فرایند اجرا شده در دنیای واقعی را داشته و قابل بهره‌برداری توسطابزارهای فرایندکاوی است.

ارزیابی چرخه آماده‌سازی داده برای فرایندکاوی با مطالعه موردی

به منظور ارزیابی چرخه فوق، فرایند کنترل کیفیت درخواست‌های کارت هوشمند ملی ایران که در سازمان ثبت احوال کشور اجرا می‌شود، برگزیده شد و عملیات آماده‌سازی داده، متناسب با چرخ ارائه شده در این مقاله انجام شد. فرایند منتخب، یکی از فرایندهای برنامه

ISSN: 2745-3673
http://www.Rcsj.ir

صدور کارت هوشمند ملی این سازمان است که در راستای ماموریت اصلی آن که همانا "اتقان[۲۰] اسناد هویتی و تابعیتی ایرانیان" است. انجام می‌شود.

در این بخش با کمک چرخه پیشنهادی، نگاره رویداد فرایند منتخب جهت انجام پروژه فرایند کاوی، آماده‌سازی می‌شود.

درک داده: داده‌های مورد نیاز برای پروژه فرایندکاوی، چرخه حیات نمونه‌های آن فرایند است که یک نمونه را (بعنوان ورودی فرایند) تا حصول نتیجه (که خروجی آن فرایند است) دنبال می‌کند. بنابراین لازم است برای تمام نمونه‌ها، کلیه رویدادهایی که انجام شده به همراه زمان و منبع آن جستجو و استخراج شوند. در این راستا پس از مصاحبه با متصدیان تولید و اجرای فرایند منتخب، این داده‌ها شناسایی شده و نگاره فرایند با استخراج ویژگی‌های مورد نیاز برای پروژه فرایندکاوی تولید شد. این ویژگی‌ها شامل نمونه، رویداد، فعالیت، زمان انجام فعالیت و منبع انجام دهنده فعالیت بود. شکل (۳) نمونه‌ای از نگاره رویداد فرایند منتخب را قبل از آماده‌سازی نشان می‌دهد.

LOG_NO	REQ_NO	ACTION	STATE	EXTRA DESC	IMAGE STATE	IMSACTION	JAAL	JAAL_DESC	VIEW BYADMIN	LOG_DATE	LOG_TIME	REGOFFICE	NINOFFICE	USER
59301429	1614226	0	6			0	1	name_err	1	24_12_2016	24_12_2016 01:12:53 PM	0		lev11
59300139	1614226	0	6			0	1	no_mach_photo	0	24_12_2016	24_12_2016 01:12:23 PM	0		lev11
55464317	15769428	0	6			0	1	shn_photo_err	0	01_12_2016	01_12_2016 08:12:02 AM	5813	5813	69IT4
57405799	15769428	4	3			0	0		0	15_12_2016	15_12_2016 12:12:40 AM	5813	5813	JIPT5
57405567	15769428	4	3			0	0		0	15_12_2016	15_12_2016 12:12:40 AM	5813	5813	JIPT5
1.55E+08	15771557	5	8			1	0		0	14_01_2018	14_01_2018 09:01:00 AM	1707	6108	CSDJ3
1.55E+08	15771557	5	4			1	0		0	14_01_2018	14_01_2018 08:01:59 AM	1707	6108	IGZN6
59933867	15771557	0	6			0	1	shn_photo_err	1	27_12_2016	27_12_2016 12:12:46 AM	1707	6108	EZRD2
90263673	15771557	0	10			0	1	name_err	0	02_05_2017	02_05_2017 09:05:28 AM	1707	6108	CSDJ4
55623933	15772113	5	6			1	1	no_mach_photo	0	03_12_2016	03_12_2016 01:12:19 PM	2581	6218	NHJM7
55869133	15772113	5	10			1	1	big_err	0	05_12_2016	05_12_2016 09:12:02 AM	2581	6218	2KF47
55592401	15773184	0	10			2	1	photo_err	0	03_12_2016	03_12_2016 10:12:04 AM	5816	3122	4233T
55478803	15773184	0	6			2	1	big_err	1	01_12_2016	01_12_2016 10:12:07 AM	5816	3122	P4X4T
52990055	15773311	0	6			0	1	photo_err	0	13_11_2016	13_11_2016 12:11:20 AM	4813	5816	69IT6
90557445	15773311	4	3			0	0		0	03_05_2017	03_05_2017 11:05:28 AM	4813	5816	JIPT5
فیلد اضافی	مورد	فیلد اضافی	فعالیت	فیلدهای اضافی						زمان اجرا		منبع		

شکل (۳): بخشی از نگاره رویداد فرایند منتخب قبل از آماده‌سازی داده

• **پاک‌سازی داده**: به جهت پاک‌سازی داده‌های نگاره تولید شده، صحت و سازگاری داده‌ها بررسی شد. با توجه به وجود سیستم تعریف شناسه مستحکمی که در سازمان ثبت احوال وجود دارد، کلیه داده‌ها از ابتدا با فرمتی مشخص و در جداولی معین ذخیره شده بودند؛ بنابراین همگون و سازگار بودند. به همین دلیل داده تکراری نیز وجود نداشت. به علاوه صحت داده‌ها مورد تایید مالک فرایند بود. زیرا کاربران دسترسی به نگاره‌ها نداشتند. داده‌ها تنها توسط سامانه ثبت می‌شد. سپس مشکلات مربوط به داده‌های گم شده، نویزها، داده‌های پرت و ... بررسی و تحلیل شد. با توجه به اهمیت بالای وضعیت‌هایی که به صورت استثنا و با فرکانس بسیار کم اتفاق می‌افتند؛ امکان حذف داده‌های پرت یا نویزها وجود نداشت. زیرا این داده‌ها در فرایند اصلی به شدت تاثیرگذارند. مانند درخواست‌های مشکوک که ۰٫۰۲٪ از کل درخواست رسیده به سازمان (تا زمان انجام این مطالعه موردی) بودند و مهمترین وضعیت در این فرایند به شمار می‌روند. این نگاره به لحاظ داده‌های گم شده نیز بررسی شد و نمونه‌های ناکامل حذف شدند.

• **تعیین سطح انتزاع داده**: با بررسی‌های انجام شده در مجموعه داده‌هایی که از این فرایند نگهداری شده بود؛ مشاهده شد، هر فعالیتی که کاربران روی درخواست‌ها انجام داده‌اند؛ با ذکر شناسه کاربری، زمان انجام فعالیت، شناسه مکان انجام فعالیت و اطلاعات خاص

[۲۰] محکم و استوار کردن

ISSN: 2745-3673
http://www.Rcsj.ir

مربوط به درخواست، توسط سامانه ثبت شده است. بنابراین نگاره این فرایند، از سطح انتزاع خوبی برخوردار بود و کلیه داده‌های مورد نیاز برای فرایندکاوی را شامل می‌شود.

• **یکپارچه‌سازی داده**: درصورت وجود منابع داده‌ای گوناگون نیاز به یکپارچه‌سازی داده‌ها وجود دارد تا تضاد و افزونگی داده‌ها برطرف گردد. اما در فرایند منتخب، کلیه داده‌های مربوط به فرایند، با نظم خاصی در یک بانک اطلاعاتی، به صورت یکپارچه ثبت و نگهداری شده بود. به همین دلیل تضاد و افزونگی داده مشاهده نشد.

تبدیل داده: عملیاتی همچون نرمال‌سازی، تغییر و تجمیع داده‌ها در این گام انجام می‌شود. از آنجایی که بانک اطلاعات فرایند منتخب یکپارچه بود، نیازی به تغییر و تجمیع داده‌ها نبود. به منظور نرمال‌سازی داده‌ها، داده‌های گم شده و ناقص حذف شدند. اما نویزها مربوط به مهم‌ترین رویداد فرایند (درخواست‌های مشکوک) بودند. بنابراین حذف نویزها به دلیل از دست دادن داده‌های خاص و با ارزش، نباید انجام می‌شد.

بررسی رانش‌های مفهومی: در جلساتی که با متصدیان تولید و اجرای این فرایند برگزار شد مشخص شد؛ این سامانه چندین بار رانش مفهومی داشته است که به صورت دستورالعمل اجرایی به کاربران ابلاغ شده و یا گاهی تغییرات کوچکی در نرم‌افزار ایجاد شده است. اما متاسفانه مستندی مبنی بر این رانش‌های مفهومی وجود نداشت. بنابراین امکان تفکیک داده‌ها در بازه‌هایی که روند اجرای فرایند ثابت بوده است، وجود ندارد.

کاهش داده: به دلیل حجم بالای درخواست‌ها (چندین میلیون درخواست) و محدودیت‌های ابزارهای موجود، امکان استفاده از کل داده‌ها وجود ندارد. بنابراین لازم است بخشی از نگاره را انتخاب نمود. اما با توجه به این که برای فرایندکاوی، ویژگی‌های خاصی از تمامی رویدادهای یک نمونه مورد نیاز است، پس باید نمونه برداری به گونه‌ای انجام گیرد که چرخه حیات نمونه‌ها کامل باشد. بدین منظور چند نوع نمونه برداری بررسی شد.

اولین نمونه این بود که داده‌ها به تفکیک مناطق جغرافیایی انتخاب شوند. مشاهده شد که این نمونه برداری صحیح نیست. زیرا کاربران و افراد مقیم مناطق مختلف کشور به دلایل گوناگونی (مانند فرهنگ، قومیت، نزدیکی به مرزها و ...) رفتارهای متفاوتی دارند.

دومین نمونه انتخاب کل نگاره در یک بازه زمانی کوتاه بود. این روش نیز نمونه خوبی از نگاره ایجاد نکرد. زیرا مشاهده شد درخواست‌های مشکوک (مهم‌ترین درخواست‌ها)، بسیار ناچیز و مانند یک نویز کنار دیگر داده‌ها بودند.

سومین نمونه انتخاب درخواست‌هایی بود که کاربر اول، وضعیت مشکوک را ثبت نموده و تصمیم‌گیری را به کاربران سطوح بالاتر واگذار کرده است. این انتخاب به دلیل اهمیت کشف درخواست‌های مشکوک، که درحقیقت علت اصلی طراحی این فرایند بود، می‌توانست انتخاب صحیحی باشد.

بنابراین به منظور کاهش داده، سومین نوع نمونه برداری انجام شد و نگاره نهایی، تولید شد.

کشف رویدادهای اجرا نشده: این فرایند ۱۵ رویداد داشت که ۱۳ رویداد آن مربوط به نمونه برداری انجام شده بود. اما در نگاره نهایی، ۱۱ رویداد مشاهده شد. بدین ترتیب رویدادهایی که هرگز اجرا نشده بودند، مشخص شدند. واضح است که این رویدادها در مدل کشف شده از نگاره نهایی نیز وجود نخواهد داشت.

در این نقطه نگاره نهایی، برای پروژه فرایندکاوی آماده شده است. حال زمان آن رسیده که یک ابزار کاوش فرایند (مانند ProM یا DISCO) انتخاب شده و پروژه فرایندکاوی اجرای شود. جدول (۱) خلاصه مراحل انجام شده در این مطالعه موردی را ارائه می‌دهد.

جدول (۱) خلاصه مطالعات موردی انجام شده

مرحله	فعالیت‌ها	نتیجه
درک داده	یافتن نمونه‌های فرایند از میان داده‌های موجود یافتن کلیه رویدادهای مربوط به هر نمونه استخراج ویژگی‌های مورد نیاز فرایندکاوی برای هر رویداد به همراه زمان و منبع مجری آن تهیه نگاره رویداد با داده‌های کشف شده	شناخت داده ساخت نگاره رویداد فرایند
پاک‌سازی داده	بررسی صحت و سازگاری داده‌های نگاره حذف افزونگی داده در نگاره تهیه شده بررسی مشکلات مربوط به داده‌های گم شده، نویزها، داده‌های پرت و ...	دست‌یابی به داده‌های صحیح و سازگار نویزها داده‌های با ارزش فرایند هستند و نباید حذف شوند. (درخواست‌های مشکوک، ۰٫۰۲٪ از کل درخواست‌ها بودند.)

ISSN: 2745-3673
http://www.Rcsj.ir

تعیین سطح انتزاع داده	تعیین سطح انتزاع داده‌های موجود در نگاره، به منظور تعیین سطح انتزاع مدل کشف شده از این داده‌ها	سطح انتزاع نگاره پایین (تمامی رویدادها، توسط سامانه در نگاره رویداد ثبت شده بود.)
یکپارچه‌سازی داده	یکپارچه‌سازی داده‌ها، جهت رفع تضاد و افزونگی داده‌های موجود در منابع داده‌ای گوناگون	عدم مشاهده تضاد و افزونگی (کلیه داده‌های مربوط به فرایند در یک جدول، به صورت یکپارچه و منظم ثبت و نگهداری شده بود.)
تبدیل داده	انجام عملیات نرمال‌سازی، تغییر و تجمیع داده‌ها	عدم نیازی به تغییر و تجمیع داده‌ها (بانک اطلاعات فرایند منتخب یکپارچه و همگون بود.) نرمال‌سازی، با حذف داده‌های ناقص و گم شده
بررسی رانش‌های مفهومی	بررسی تغییرات روند اجرای فرایند در حال اجرا	رانش مفهومی اتفاق افتاده بود. اما مستند نشده بود.
کاهش داده	نمونه برداری با حفظ چرخه حیات نمونه‌ها	انتخاب درخواست‌هایی با یک رویداد مشکوک
کشف رویدادهای اجرا نشده	- مقایسه لیست رویدادهای فرایند مربوط به نمونه برداری انجام شده، با لیست رویدادهای موجود در نگاره نمونه برداری شده	یافتن ۲ رویداد هرگز اجرا نشده

شکل (۴) بخشی از نگاره رویداد فرایند منتخب را پس از آماده سازی برای فرایندکاوی نشان می‌دهد.

REQ_NO	STATE	TIMESTAMP	REQOFFICE	NINOFFICE	USER
44701855	6	26-04-2018 10:04:37	3131	3128	4A048
44701855	10	26-04-2018 11:04:42	3131	3128	5B264
44706916	6	25-04-2018 06:04:58	4149	5147	RHCSK
44706916	3	29-04-2018 06:04:24	4149	5147	12D80
44718005	3	08-05-2018 08:05:43	8157	7148	6E1QC
44718005	6	26-04-2018 09:04:51	8157	7148	QR7EC
44724921	10	16-08-2018 09:08:41	2128	6149	67C6C
44724921	6	26-04-2018 07:04:50	2128	6149	X8OJ8
44734870	6	01-05-2018 08:05:35	4171	4156	X9RND
44734870	3	01-05-2018 11:05:03	4171	4156	J99LE
44736360	6	12-06-2018 06:06:38	3152	5155	M8DSR
44740477	3	07-05-2018 07:05:41	3153	5147	R8HSK
44740477	6	26-04-2018 08:04:19	3153	5147	19280
44741722	6	26-04-2018 08:04:29	1167	1167	Y45YQ
44741722	3	28-04-2018 01:04:46	1167	1167	56037
مورد	فعالیت	زمان اجرا	منبع		

شکل (۴): بخشی از نگاره آماده‌سازی شده برای پروژه فرایندکاوی

نتیجه گیری

در این مقاله ترتیبی از مراحل آماده‌سازی داده برای انجام پروژه‌های فرایندکاوی، پیشنهادشده است. بدین منظور ابتدا مراحل آماده‌سازی داده برای پردازش‌های مبتنی بر داده مانند داده‌کاوی، وب‌کاوی، متن‌کاوی و ... بررسی شد و مراحل اصلی آن که عبارتند از درک داده، پاک‌سازی داده، یکپارچه‌سازی داده، تبدیل داده و کاهش داده، توصیف شد. سپس چالش‌های موجود در زمینه آماده‌سازی داده برای فرایندکاوی، مورد بررسی و تحلیل قرار گرفت و بر مبنای این چالش‌ها چند مرحله به مراحل فوق اضافه شد. این مقاله پیشنهاد می‌دهد، به منظور آماده‌سازی داده برای یک پروژه فرایندکاوی، به ترتیب مراحل زیر طی شود:

۱- درک داده ۲- پاک‌سازی داده ۳- تعیین سطح انتزاع داده ۴- یکپارچه‌سازی داده ۵- تبدیل داده ۶- بررسی رانش‌های مفهومی ۷- کاهش داده ۸- کشف رویدادهای اجرا نشده

ISSN: 2745-3673
http://www.Rcsj.ir

مراحل اضافه شده به مراحل اصلی آماده سازی داده عبارتند از:

- تعیین سطح انتزاع داده، که بر سطح انتزاع مدل کشف شده تاثیر مستقیم دارد.
- بررسی رانش‌های مفهومی، که بر کنترل جریان مدل کشف شده تاثیرگزار است.
- کشف رویدادهای اجرا نشده، زیرا این رویدادها در مدل کشف شده نیز وجود نخواهند داشت.

در نهایت این چرخه با کمک یک مطالعه موردی ارزیابی شد. بدین منظور فرایند کنترل کیفیت درخواست‌های کارت هوشمند ملی ایران، انتخاب شده و یک نگاره رویداد برای انجام پروژه فرایندکاوی، آماده سازی شد.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[۱] آسمان منظر سارا، " طبقه بندی راهکارهای پیش پردازش داده در داده کاوی و کشف دانش"، هفتمین کنفرانس ملی و اولین کنفرانس بین المللی مدیریت دانش، تهران، موسسه اطلاع رسانی نفت، گاز و پتروشیمی، ۱۳۹۴.

[۲] خدا مرادی, محمد و پریسا دهقانی، "بررسی تکنیک‌های آماده‌سازی داده‌ها در داده‌کاوی و ارائه راهکار بهینه پیش پردازش در هر حوضه‌های کاربردی داده‌کاوی، کنفرانس ملی علوم مهندسی، ایده‌های نو (۸)، تنکابن، موسسه آموزش عالی آیندگان تنکابن، ۱۳۹۳.

[۳] متولی‌حقی سیدمحمد، احمدیان طبسی حمیدرضا، سجادی سیدناصر، "آماده‌سازی داده‌ها برای داده‌کاوی تکنیک‌ها و کاربرد آن"، اولین همایش ملی فناوری اطلاعات و شبکه‌های کامپیوتری دانشگاه پیام نور- واحد طبس- ۲۵ بهمن ۱۳۹۱.

[۴] کرمی راضیه، ملک‌جعفریان ملیحه‌سادات، "اهمیت پردازش داده‌ها"، مجله آمار، شماره ۸، صفحه ۳۴-۳۶، مهر و آبان ۱۳۹۳.

[۵] بابائی مریم، صفاریزدی زهرا، سرایی محمدحسین، "معرفی و مقایسه روش‌های پیش‌پردازش داده برای کاربردهای مختلف داده کاوی"، دومین کنفرانس داده کاوی ایران، ۱۳۸۷.

[6] Alexandropoulos S.-A. N., Kotsiantis S. B., Vrahatis M. N., "Data preprocessing in predictive data mining", The Knowledge Engineering Review, Vol. 34, e1: 1–33. © Cambridge University Press, 2019.

[7] Ram´ırez-Gallego Sergio et al., "A survey on Data Preprocessing for Data Stream Mining: Current status and future directions", Neurocomputing, Elsevier, Volume 239, 24 May 2017, Pages 39-57.

[8] Alasadi S. A., Bhaya W. S., "Review of Data Preprocessing Techniques in Data Mining", Journal of Engineering and Applied Sciences 12(16): 4102-4107,2017.

[9] Yuxin Chen et al., "SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data",GigaScience, 7, 2018, 1–6.

[10] van der Aalst Wil et al,ترجمه: آصف پورمعصومی , "process_mining_manifesto", http://www.win.tue.nl/ieeetfpm/lib/exe/fetch.php?media=shared:pmm-persian-v1.pdf

[11] Danubianu Mirela, "Step by step data preprocessing for data mining. Case study", Conference, InfoTech-2015, Bulgaria.

[12] Zhang Chengqi, Zhang Shichao, Yang Qiang, "Data preparation for data mining", Taylor & Francis , 17:375-381, 2003.

[13] Wen L.J et al., "A Novel Approach for Process Mining Based on Event Types" , 2009, Journal of Intelligent Information System, 32(2), 163-190.

[14] Wen L.J et al.," Mining process models with prime invisible tasks.", Data and knowledge Engineering 2010, 69(10), 999-1021.

[15] Wang J.M., Wen L.J., "Discovering Process Knowledge from Event Logs", Communications of the CCF, 2012, 8 (6), 63-68.

[16] Salas H.A et al., "A Spatial-based KDD Process to Better Understand the Spatiotemporal Phenomena" ,25th International Conference on Advanced Information Systems Engineering, Jun 2013, Valencia, Spain.

[17] Shrikanth Narayanan et al., "Integration and Automation of Data Preparation and Data Mining", 2014 IEEE International Conference Data Mining Workshop

[18] Bogorny Vania, Martins Engel Paulo, Otavio Alvares Luis, "Spatial Data Preparation for Knowledge Discovery", IEEE Computer Graphics pp. 24 (5), 8, 2005.

[19] Van der Aalst W.M.P., "Process Mining: Data Science in Action", second edition, Springer, Berlin (2016)

[20] Van der Aalst W.M.P., "Process Mining: Discovery, Conformance and Enhancement of Business Processes". Springer, Berlin (2011)

[21] van der Aalst W.M.P., "Process Mining: A historical perspective", Process Mining Camp 2013 – Fluxicon.

ISSN: 2745-3673
http://www.Rcsj.ir

[22] ZooniAtuba, Vidyavati B.M., "A Survey on Process Mining", WIREs Data Mining and Knowledge Discovery, Volume 8, 2018, page: 1348-1351.
[23] Park Sungbum, Kang Young Sik, "A Study of Process Mining-based Business Process Innovation", ITQM 2016,ELSEVIER, Procedia Computer Science, volume 19, pp: 734-743.
[24] Ceravolo Paolo et al, "Translating Process Mining Results into Intelligible Business Information", 11thInternational Knowledge Management Coference, Hagen, Germany, 2016
[25] Rahm E., Do H.H., "Data Cleaning: Problems and Current Approaches", IEEE Bulletin on Data Engineering 23:4, 3-13, 2000
[26] Minker J., "Logic Based Artificial Intelligence", Springer US, 2000
[27] Chapman P. & Kerber R. (NCR), Clinton J. & Khabaza T. & Shearer C. (SPSS), Reinartz T. & Rüdiger W. (Daimler Chrysler), "CRISP-DM 1.0 step-by-step data mining guide", copyright 1999-2000
[28] Buijs J.C.A.M., van Dongen B.F. and van der Aalst W.M.P., "Mining Configurable Process Models from Collections of Event Logs", In Proceedings of the 11thInternational Conference on Business Process Management, 2013, 33-48.
[29] Jagadeesh Chandra Bose R.P., van der Aalst W.M.P., ZliobaiteI and Pechenizkiy M., "Dealing with Concept Drifts in Process Mining", IEEE Transactions on Neural Networks and Learning Systems, 25(1), 154-171, 2014
[30] Leemans S.J. J., D. Fahland, van der Aalst W.M.P., "Discovering block-structured process models from incomplete event logs", In Proceedings of the International Conference on Applications and Theory of Petri Nets and Concurrency, 2014, 91–110
[31] Shugurov I.S., Mitsyuk A.A., "Generation of a Set of Event Logs with Noise", 8thSpring/Summer Young Researchers Colloquium on Software Engineering, Russia, 2014