

شناسایی رانش مفهومی در نگاره‌های فرایند کسب و کار با استفاده از یادگیری عمیق

فاطمه خجسته^{*}، محسن کاهانی و بهشید بهکمال

۱۳۹۷/۰۶/۱۱ تاریخ پذیرش: ۱۳۹۸/۰۶/۱۱ تاریخ انتشار: ۱۳۹۹/۱۲/۰۴ نوع مطالعه: پژوهشی

گروه رایانه دانشگاه فردوسی، مشهد، ایران

چکیده

فرایندهای کسب و کار در دنیای واقعی بسیار پیچیده هستند و متناسب با تحولات محیطی دچار تغییر می‌شوند. این در حالی است که روش‌های کشف فرایند پایه، قادر به شناسایی این تغییرات نیستند و تنها فرایندهای ثابت را تحلیل می‌کنند؛ ازین‌رو، روش‌هایی بهمنظور شناسایی رانش مفهومی در فرایندهای کسب و کار مطرح شدند. همه روش‌های موجود در این حوزه، با انتخاب ویژگی‌ها و مقایسه آنها با استفاده از پنجره سعی در شناسایی این تغییرات دارد. انتخاب ویژگی مناسب و همچنین اندازه مناسب پنجره چالش‌های اصلی این روش‌ها به شمار می‌آیند. در این پژوهش، با بیان مفهوم تعییه واژه در دنیای پردازش زبان طبیعی است، روشنی خودکار و مستقل از پنجره بهمنظور شناسایی رانش ناگهانی در نگاره‌های کسب و کار ارائه کرده‌ایم. استفاده از روش تعییه دنباله، این امکان را فراهم می‌کند که انواع روابط میان دنباله‌ها و رویدادها را استخراج و رانش‌های موجود در فرایندها را شناسایی کنیم. ارزیابی‌ها نشان می‌دهد که روش پیشنهادی نسبت به روش‌های موجود دقت بالاتر و تأخیر شناسایی رانش کمتری دارد.

وازگان کلیدی: فرایندکاوی، رانش مفهومی، تغییرات فرایند، تعییه واژه

Concept drift detection in business process logs using deep learning

Fatemeh Khojasteh*, Mohsen Kahani & Behshis Behkamal

Departeman Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

Abstract

Process mining provides a bridge between process modeling and analysis on the one hand and data mining on the other hand. Process mining aims at discovering, monitoring, and improving real processes by extracting knowledge from event logs. However, as most business processes change over time (e.g. the effects of new legislation, seasonal effects and etc.), traditional process mining techniques cannot capture such “second-order dynamics” and analyze these processes as if they are in steady-state. Such changes can significantly impact the performance of processes. Hence, for the process management, it is crucial that changes in processes be discovered and analyzed. Process change detection is also known as business process drift detection.

All the existing methods for process drift detection are dependent on the size of windows used for detecting changes. Identifying convenient features that characterize the relations between traces or events is another challenge in most methods. In this thesis, we propose an automated and window-independent approach for detecting sudden business process drifts by introducing the notion of trace embedding. Using trace embedding makes it possible to automatically extract all features from the relations between traces.

* Corresponding author

*نویسنده عهده‌دار مکاتبات

سال ۱۳۹۹ شماره ۴ پیاپی ۴۶

• تاریخ ارسال مقاله: ۱۳۹۷/۰۷/۱۸ • تاریخ پذیرش: ۱۳۹۸/۰۶/۱۱ • تاریخ انتشار: ۱۳۹۹/۱۲/۰۴ • نوع مطالعه: پژوهشی

We show that the proposed approach outperforms all the existing methods in respect of its significantly higher accuracy and lower detection delay.

Keywords: process mining, concept drifts, process changes, word embedding

روش‌های متنوعی را جهت شناسایی تغییرات فرایند ارائه دهند.

در سال‌های اخیر روشنگرهای متعددی برای شناسایی رانش مفهومی در نگاره‌های کسب‌وکار مطرح شده است که هر کدام بر روی دیدگاه (ها) و رانش‌های مختلفی تمرکز کرده‌اند. بیشتر روشنگرهای موجود نیازمند به استخراج ویژگی‌ها (ها) از دنباله^{۱۲}ها و رویدادها هستند. نوع ویژگی‌ها (های) انتخاب شده بر دقت روشنگرها تأثیرگذار است. از طرفی، همه روشنگرهای موجود از دو پنجه متواالی استفاده کرده و مقادیر ویژگی‌ها در این دو پنجه را به روشنگرهای مختلف مورد مقایسه قرار می‌دهند. نتایج ارائه شده در این روشنگرها نشان می‌دهد که دقت شناسایی رانش‌ها بهشت است به نوع و اندازه پنجه استفاده شده وابسته است؛ بنابراین بهنظر می‌رسد که می‌توان با ارائه روشنگرها مستقل از پنجه که توانایی استخراج ویژگی‌ها به صورت خودکار را داشته باشد، چالش‌های پیشین را برطرف کرد. ایده پیشنهادی بدین صورت است که با توجه به نوع نگاره ورودی و با به کارگیری مفهوم تعییه دنباله^{۱۳}، می‌توان دنباله‌های موجود در نگاره کسب‌وکار را به بردارهایی مدل کرد؛ به طوری که روابط میان بردارها در فضای برداری، بیان‌گر روابط میان دنباله‌ها در نگاره هستند؛ سپس می‌توان با استفاده از روشنگر خوشبندی، دنباله‌های هم‌رخداد را شناسایی کرده و نحوه توزیع آن‌ها را تعیین کرد؛ درنهایت، با استفاده از تبدیل فوریه و شناسایی تغییرات اصلی قادر خواهیم بود رانش‌ها و مکان دقیق آن‌ها را شناسایی کنیم.

نتایج ارزیابی حاکی از این است که روشنگرها بهبود قابل توجهی در معیارهای امتیاز F- و متوسط تأخیر^{۱۴} نسبت به سایر روشنگرهای ایجاد کرده است؛ بنابراین، نوآوری‌های این پژوهش به صورت زیر خلاصه می‌شود:

- ارائه ایده تعییه دنباله به منظور استخراج خودکار روابط میان دنباله‌ها و رویدادها

- حذف پنجه به منظور شناسایی رانش‌ها و استفاده از معیار شباهت دنباله‌ها
- استفاده از تبدیل فوریه به منظور شناسایی مکان دقیق رانش و کاهش متوسط تأخیر

¹² Trace

¹³ Trace embedding

¹⁴ Mean delay

۱- مقدمه

فرایندکاوی^۱ یک زمینه پژوهشی به نسبه جدید است که با استخراج دانش از نگاره‌های رویداد^۲، سعی در کشف^۳، بررسی انطباق^۴ و بهبود^۵ فرایندکاوی واقعی دارد [۱]. دو دلیل اصلی برای افزایش پژوهش‌ها در حوزه فرایندکاوی وجود دارد. از یکسو، به کمک رویدادهای ثبت شده در نگاره رویداد می‌توان اطلاعات ارزشمندی درباره تاریخچه فرایندکاوی سازمانی به دست آورد. از سوی دیگر، با توجه به فضای رقابتی کسب‌وکار و تغییرات محیطی، نیاز به بهبود و پشتیبانی از فرایندکاوی کسب‌وکار اهمیت بسیاری دارد.

فرایندکاوی کسب‌وکار^۶ را می‌توان از دیدگاه‌های^۷ مختلف مانند کنترل جریان^۸، داده و منبع^۹ مورد تحلیل قرار داد [۲]. فرایندکاوی کسب‌وکار در دنیای واقعی پویا هستند؛ به گونه‌ای که در زمان‌های مختلف، گونه‌های متفاوتی از یک فرایند اجرا می‌شود. وجود تغییرات بسیار در نرخ عرضه و تقاضا، تغییرات فصلی، وضع قوانین جدید، بلایای طبیعی و غیره از جمله واقعی هستند که سازمان‌ها را مجبور می‌سازد تا فرایندکاوی خود را تغییر دهند. این در حالی است که پیش‌تر روشنگرهای کشف فرایند، برای فرایندکاوی پایدار و ساخت‌یافته طراحی شده‌اند و مدل حاصل از آن‌ها برای فرایندکاوی دنیای واقعی یک مدل اسپاگتی شکل است. از طرفی دیگر، سازمان‌ها در تلاش هستند با ساده‌سازی فرایندکاوی کسب‌وکار، هزینه‌ها را کاهش داده و کارایی را بهبود بخشند. بنابراین، شناسایی تغییرات منتظره یا غیرمنتظره در رفتار فرایندکاوی و تحلیل آن‌ها، اطلاعات ارزشمندی را برای ذی‌نفعان آشکار می‌سازد. رانش مفهومی^{۱۱} زمانی رخ می‌دهد که فرایند در هنگام تحلیل دچار تغییر شده باشد [۱]. نیاز به مدیریت این تغییرات، پژوهش‌گران در حوزه فرایندکاوی را بر آن داشت تا

¹ Process mining

² Event log

³ Discovery

⁴ Conformance checking

⁵ Enhancement

⁶ Business process

⁷ Prospective

⁸ Control-flow

⁹ Resource

¹⁰ Variant

¹¹ Concept drift

فصلنامه

پژوهش‌های
برداری و روزانه

مختلفی هستند. برای هر نمونه $c \in C$ و هر نام $n \in AN$ $\#_n(c)$ نشان‌دهنده مقدار صفت n برای رویداد e است [3]. هر نمونه دارای یک صفت اجباری مخصوص به نام دنباله است: $E^* \ni c \#_{trace}(c)$. یک دنباله یک مجموعه محدود از رویدادها است، $\sigma \in E^*$.

نگاره رویداد یک مجموعه از نمونه‌های است، $L \subseteq C$.

۲-۲- رانش مفهومی

منظور از رانش مفهومی در یادگیری ماشین و داده‌کاوی، ایجاد تغییر پیش‌بینی‌شده در رابطه بین داده ورودی و متغیر هدف در طول زمان است [4] یا به عبارت دیگر هر تغییر در نحوه توزیع داده در طی زمان را رانش مفهومی گویند.

علاوه‌بر حوزه‌های یادگیری ماشین و داده‌کاوی، در سال‌های اخیر شناسایی رانش مفهومی در حوزه فرایندکاوی نیز مورد توجه و پژوهش قرار گرفته است. فرایندهای کسب و کار دنیای واقعی پویا هستند و در بازه‌های زمانی مختلف، گونه‌های اجرایی متفاوتی از یک فرایند اجرا می‌شوند. این در حالی است که روش‌های فرایندکاوی پایه، برای فرایندهای پایدار و ساخت‌یافته طراحی شده‌اند؛ بنابراین، مدل حاصل از این روش‌ها برای فرایندهای دنیای واقعی یک مدل اسپاگتی شکل است. مدل‌های اسپاگتی شکل، بسیار پیچیده هستند و به همین دلیل توصیف دقیق و واضحی از فرایند اصلی در حال اجرا ارائه نمی‌دهند؛ بنابراین، فرض ثابت بودن فرایند در حین اجرا، یک فرض غیرواقعی است و شناسایی تغییرات قابل‌انتظار و غیرقابل‌انتظار در رفتار فرایندها و تحلیل آن‌ها، اطلاعات ارزشمندی را برای ذی‌نفعان آشکار می‌سازد. در [1] از یازده موضوع به عنوان چالش‌های موجود در حوزه فرایندکاوی یاد شده که چهارمین چالش مطرح شده در آن، شناسایی رانش مفهومی در فرایندهای کسب و کار است و آن را به این صورت تعریف می‌کند:

"رانش مفهومی زمانی رخ می‌دهد که فرایند در هنگام تحلیل دچار تغییر شده باشد [1]."

همچنین، با توجه به آنچه در [5] گفته شده است، مسئله شناسایی رانش مفهومی در فرایندهای کسب و کار از دیدگاه آماری عبارت است از شناسایی نقطه‌ای از زمان که میان رفتار فرایند در قبل و بعد از این نقطه، اختلاف قابل‌مالحظه آماری وجود داشته باشد.

ادامه مقاله بدین صورت بخش‌بندی شده است. در بخش دوم، مرور کارهای گذشته مطرح می‌شود. در ابتدای این بخش به بیان مفاهیم اولیه پرداخته شده و سپس روش‌های موجود در شناسایی رانش فرایند بحث شده است. روش پیشنهادی در بخش سوم بیان شده است. بخش چهارم در بردازندۀ نتایج و ارزیابی‌های انجام‌شده بر روی نگاره‌های ساختگی و واقعی است، درنهایت، در بخش پنجم، نتیجه‌گیری و کارهای آینده بیان می‌شود.

۲- مرور کارهای گذشته

هدف از این بخش، بیان مفاهیم اولیه در حوزه فرایندکاوی، مفهوم رانش مفهومی و همچنین روش‌های مختلفی است که تاکنون در زمینه شناسایی رانش مفهومی در فرایندهای کسب و کار ارائه شده‌اند.

۲-۱- مفاهیم اولیه

نقشه شروع برای فرایندکاوی، وجود یک نگاره رویداد است. تمامی روش‌های فرایندکاوی بر این فرض هستند که امکان ثبت ترتیبی رویدادها به گونه‌ای که هر رویداد به یک فعالیت و به یک نمونه (مورد) خاص اشاره کند، وجود دارد. به عبارت دیگر، رویدادها به نمونه‌های فرایند مرتبط و از طریق نام فعالیتها توصیف می‌شوند. رویدادها در یک نمونه فرایند مرتب هستند؛ بنابراین، یک نمونه فرایند، اغلب به صورت یک دنباله بر روی فعالیتها نشان داده می‌شود؛ به علاوه، هر رویداد می‌تواند صفاتی مانند منبع (شخص یا دستگاه) آغاز‌کننده فعالیت، برچسب زمانی رویداد، صفات داده‌ای (مقدار یا نوع مشتری) داشته باشد [3].

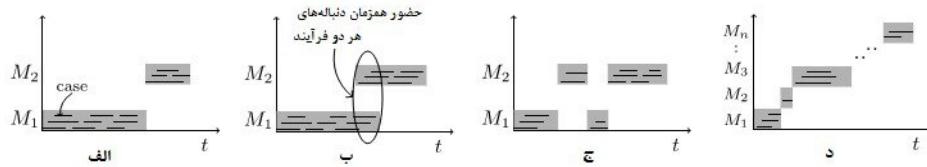
(رویداد و صفت): فرض کنید E مجموعه‌ای از رویداد‌ها باشد. رویدادها با صفات متفاوتی مشخص می‌شوند. به عنوان مثال، هر رویداد می‌تواند یک برچسب زمانی، نام فعالیت^۱، نام منبع، هزینه و غیره داشته باشد. همچنین، فرض کنید AN مجموعه‌ای از نام صفت‌های مختلف باشد. برای هر رویداد $e \in E$ و نام $n \in AN$ $\#_n(e)$ نشان‌دهنده مقدار صفت n برای رویداد e است. اگر رویداد e ، صفتی با نام n نداشته باشد، آن‌گاه $\#_n(e) = \perp$ خواهد بود [3].

(نمونه، دنباله، نگاره رویداد): فرض کنید C مجموعه‌ای از نمونه‌ها است. نمونه‌ها نیز مانند رویدادها دارای صفات

¹ Event

² Activity

³ Casc

(شکل-۱): انواع رانش. الف) رانش ناگهانی، ب) رانش تدریجی، ج) رانش متناوب، د) رانش افزایشی.^[۱]

(Figure-1): Different types of drifts. (a) Sudden drift. (b) Gradual drift. (c) Recurring drift. (d) Incremental drift

وجود دارد. به عنوان مثال، سازمانی را در نظر بگیرید که فرایند تحويل محصولاتش را تغییر داده است. این سازمان برای مشتریانی که سفارش خود را قبل از تغییر فرایند ثبت کرده باشد از فرایند قدیمی و برای مشتریان جدید از فرایند جدید استفاده می‌کند.

• رانش متناوب^۹: مجموعه‌ای از فرایندها که بعد از مدت زمانی مجدد ظاهر می‌شوند. این تغییرات ممکن است به دلیل تغییرات فصلی به وجود آیند. به عنوان مثال، یک شرکت مسافرتی از فرایندی متفاوت در ایام عید نوروز استفاده می‌کند. همان‌طور که در شکل (۱-ج) مشاهده می‌شود فرایند M_2 در بازه‌های زمانی خاصی، جایگزین فرایند M_1 شده است.

• رانش افزایشی^{۱۰}: همان‌طور که در شکل (۱-د) مشاهده می‌شود، فرایند M_N با تغییرات کوچک و افزایشی جایگزین فرایند M_1 شده است. به عنوان مثال، می‌توان سازمانی را در نظر گرفت که برای بهبود عملکردش، تغییرات کم اما افزایشی و بلندمدت را بر روی فرایند اولیه خود اعمال می‌کند.

۲-۳-روش‌های شناسایی رانش فرایند

بیشتر روش‌های موجود، به منظور شناسایی روابط میان دنباله‌ها، نیاز به استخراج ویژگی دارند. چنانچه ویژگی‌های مناسبی انتخاب نشود، روش بیان شده توانایی شناسایی همه انواع تغییرات را نخواهد داشت و دقت آن کاهش می‌یابد؛ همچنین، همه روش‌های موجود به منظور شناسایی رانش فرایند از پنجره ثابت یا تطبیقی استفاده می‌کنند. استفاده از پنجره به این صورت است که دو پنجره متوالی بر روی جریانی از ویژگی‌های استخراج شده حرکت داده می‌شود و مقادیر آن‌ها به روش‌های گوناگون مورد مقایسه قرار می‌گیرد. اگرچه استفاده از پنجره تطبیقی، دقت را افزایش داده است، اما انتخاب اولیه اندازه پنجره، همچنان یک چالش مهم به شمار می‌آید.

^۹ Recurring drift

^{۱۰} Incremental drift

از روش‌های مطرح شده برای شناسایی رانش مفهومی در حوزه‌های یادگیری ماشین و داده‌کاوی نمی‌توان در فرایند کاوی استفاده کرد، چراکه تغییرات در این دو حوزه نسبت به فرایند کاوی، ساختار بسیار ساده‌تری دارند. از جمله تغییراتی که در مدل فرایند کسب‌وکار ممکن است رخداد، می‌توان به هم‌زمانی^۱ در مدل فرایند، حلقه‌ها^۲، انتخاب‌ها^۳ و غیره اشاره کرد که این تغییرات، ساختار بسیار پیچیده‌تری نسبت به تغییرات در داده‌کاوی و یادگیری ماشین دارند [۳]. با توجه به مدت زمانی که یک تغییر در فرایند وجود دارد، تغییرات را می‌توان در دو دسته ناپایدار^۴ و پایدار^۵ طبقه‌بندی کرد. تغییرات ناپایدار، تغییراتی هستند که عمر کوتاهی دارند و فقط بر تعداد محدودی از نمونه‌ها تأثیر می‌گذارند؛ در حالی که تغییرات پایدار، دائمی هستند و برای یک مدت طولانی باقی می‌مانند [۶]. شناسایی تغییرات ناپایدار به دلیل کمبود داده امکان‌پذیر نیست. به تغییرات ناپایدار، داده‌های پرت^۶ یا نوافه نیز گفته می‌شود. به منظور شناسایی رانش‌های مفهومی باید تغییرات دائمی را شناسایی کرد [۳] چهار نوع رانش در ادامه بیان خواهد شد.

• رانش ناگهانی^۷: در این نوع رانش، یک فرایند جدید جایگزین فرایند موجود می‌شود. به عنوان مثال، سازمانی را در نظر بگیرید که به دلیل تغییر قوانین، باید فرایند جدیدی را جایگزین فرایند قبلی اش نماید. شکل (۱-الف)، نشان‌دهنده رانش ناگهانی است که در آن، فرایند M_2 جایگزین فرایند M_1 شده است.

• رانش تدریجی^۸: در این نوع رانش، یک فرایند جدید به صورت تدریجی جایگزین فرایند موجود می‌شود. همان‌طور که در شکل (۱-ب) نشان داده شده است، برخلاف رانش ناگهانی امکان حضور هم‌زمان هر دو فرایند

^۱ Concurrency

^۲ Loops

^۳ Choices

^۴ Momentary

^۵ Permanent

^۶ Outlier

^۷ Sudden drift

^۸ Gradual drift



مارکوف است که از یک ماتریس شباهت به عنوان ورودی استفاده می‌کند. به منظور شناسایی نقاط تغییر، ماتریس شباهت به دست آمده از هر زیرنگاره در دو پنجره متولی مورد مقایسه قرار می‌گیرد. از جمله نقاط ضعف این روش، استفاده از پنجره و نیازمندی به تنظیمات دستی است.

از مفهوم بردار پاریک^۱ در [10] استفاده شده است.

این بردار برای مجموعه‌ای از دنباله‌ها (دنباله‌های یادگیری) محاسبه شده و سپس یک چندضلعی با توجه به بردارهای تشکیل شده ایجاد می‌کند. زیرتولی حاصل از هر دنباله بعدی، مدل و بررسی می‌شود که آیا در داخل چندضلعی قرار می‌گیرد یا خیر. چنانچه یک دنباله در داخل چندضلعی قرار بگیرد، یعنی متعلق به فرایند یکسان است؛ در حالی که اگر تعداد قابل توجهی از دنباله‌ها خارج از چندضلعی قرار بگیرند، نشان‌دهنده وجود تغییر در فرایند است؛ یا به عبارت دیگر یک رانش شناسایی شده است. در این روش امکان شناسایی نقطه تغییر وجود ندارد و تنها می‌تواند رانش‌های ناگهانی را شناسایی کند. همچنین نیاز به محاسبات سنگینی دارد.

برخلاف روش ارائه شده در [3]، الگوریتم مطرح شده در [11] به ساختار توجه کرده است، احتمال را مدنظر قرار می‌دهد و از اutomاتای محدود قطعی احتمالی^۲ (PDFA) استفاده می‌کند. روند کلی روش به این صورت است که زیرنگاره‌ها را به کمک دو پنجره متولی متحرک استخراج، سپس از الگوریتم آلفا استفاده کرده و شبکه پتری مربوط به هر پنجره را تولید می‌کند. با توجه به این که شبکه پتری یک شبکه غیر احتمالی است، به منظور مقایسه با توزیع مدل پایه^۳، آن را به PDFA تبدیل می‌کند. با استفاده از آزمون آماری بررسی می‌کند که آیا نمایش PDFA (نحوه توزیع) هر زیرنگاره با مدل پایه به طور قابل ملاحظه‌ای تغییر کرده است یا خیر. وقت روش به مقدار داده مورد استفاده وابسته است. از طرفی مقدار زیاد داده باعث می‌شود که زمان اجرا افزایش یابد؛ بنابراین، ایجاد یک تعادل میان دقت و زمان اجرا یکی از چالش‌های این روش است. همچنین، این روش تنها توانایی شناسایی تغییرات بزرگ را دارد.

یک روش خودکار مبتنی بر مفهوم اجرا^۴ در [12]

مطرح شده است. در این روش به منظور شناسایی رانش، برای هر دنباله یک اجرا ساخته می‌شود؛ سپس با استفاده از

در [3,5] روشی به منظور شناسایی رانش‌های ناگهانی و انواع خاصی از رانش‌ها تدریجی ارائه شده است. ویژگی‌هایی به منظور شناسایی روابط میان دنباله‌ها، توسط کاربر استخراج می‌شود. در برخی موارد، به منظور انتخاب ویژگی مناسب، کاربر نیازمند داشت قبلی درباره نکاره است؛ سپس، از آزمون آماری، به منظور مقایسه ویژگی‌ها در دو پنجره متولی استفاده می‌شود. روش ارائه شده در [3,5] دارای چند مشکل اصلی هستند. این روش به صورت خودکار عمل نمی‌کند. دقت روش به طور قابل ملاحظه‌ای تحت تاثیر اندازه پنجره انتخاب شده است. با توجه به نوع ویژگی انتخاب شده، ممکن است همه انواع رانش‌ها شناسایی نشوند. روش مطرح شده در [3]، با بیان مفهوم پنجره تطبیقی قصد در توسعه متد پیشنهادی [7] را دارد. در این روش، دو مقدار کمینه و بیشینه برای پنجره تنظیم می‌شود و اندازه کمینه پنجره تا زمانی که یک تغییر شناسایی شود یا به اندازه بیشینه برسد، افزایش می‌یابد. با توجه به این که، اندازه کمینه و بیشینه توسط کاربر تعیین می‌شود، دقت روش کاهش می‌یابد. به عبارت دیگر، اگر اندازه کمینه تعیین شده خیلی کوچک باشد، نویفه‌ها را به عنوان رانش شناسایی می‌کند و اگر اندازه بیشینه خیلی بزرگ باشد، ممکن است برخی از رانش‌ها را شناسایی نکند. [7] با در نظر گرفتن پنجره‌هایی در مقیاس زمانی، قادر به شناسایی برخی از تغییرات تدریجی است. انتخاب بازه زمانی مناسب برای پنجره، چالش بعدی این روش به منظور شناسایی رانش تدریجی است.

الگوریتم بیان شده در [8] با استفاده از پنجره ثابت، قادر به شناسایی رانش‌های ناگهانی، تدریجی و متناوب است. در این روش، دنباله‌ها بر اساس فاصله بین زوج فعالیت‌هایشان خوشبندی می‌شوند. پارامترهای مختلفی دقت این روش را تحت تأثیر قرار می‌دهند. از جمله مهم‌ترین پارامترها، اندازه پنجره استفاده شده است. همچنین، این روش برای نکاره‌هایی شامل گونه فرایندی‌های زیاد دقت مناسبی ندارد. با توجه به این که، این روش از فاصله بین زوج فعالیت‌ها به منظور شناسایی رانش فرایند استفاده می‌کند، آن دسته از رانش‌هایی را که منجر به تغییر در فاصله بین فعالیت‌ها نمی‌شوند شناسایی نمی‌کند.

روش ارائه شده در [9] تنها روشی است که هر دو دیدگاه کنترل جریان و داده را در نظر می‌گیرد. این روش از خوشبندی به منظور شناسایی تغییرات ناگهانی استفاده می‌کند. خوشبندی آن بر اساس الگوریتم خوشبندی

¹ Parikh

² Probabilistic deterministic finite automata

³ Ground truth

⁴ Run



آزمون آماری، نحوه توزیع اجرها در دو پنجره تطبیقی متولی مورد مقایسه قرار می‌گیرد. این روش تنها برای فرایندهای ساخت‌یافته طراحی شده و دقت آن بهاندازه ابتدایی پنجره وابسته است.

الگوریتم مطرح شده در [13] سعی دارد با بهبود روش موجود در [12] روش‌های تدریجی را شناسایی کند. ایده اصلی روش مطرح شده بر این اساس است که روش‌های تدریجی به شکل دو روش ناگهانی متولی ظاهر می‌شوند. بنابراین، بعد از شناسایی روش‌های ناگهانی با استفاده از روش [13]، یک آزمون آماری را بر روی سبقه‌نما^۱ اجراها اعمال می‌کنند تا مشخص کنند که روش‌های ناگهانی شناسایی شده، تغییرات جداگانه یا یک روش تدریجی هستند. از جمله محدودیت‌های این روش، تغییر اندازه پنجره به منظور تعیین یک حد آستانه مناسب بین دقت و تأخیر شناسایی روش است؛ همچنین، این روش برای نگاره‌ها با تعداد زیادی گونه متفاوت کارنمد نیست.

رونده کلی [14] به این ترتیب است که نگاره رویداد را به جریانی از دنباله‌ها تبدیل کرده و روابط میان دنباله‌ها در دو پنجره متولی را با استفاده از معیار وابستگی استخراج کرده و آن‌ها را با استفاده از ماتریس روابط بیان می‌کند. معیار وابستگی استفاده شده در این روش کاوش‌گر ابتكاری است که نسبت به دیگر معیارهای شباهت عملکرد بهتری در مواجهه با نگاره‌های ناکامل دارد. در هر مرحله، ماتریس‌های حاصل از دو پنجره را مقایسه می‌کند. اگر ماتریس دارای اختلاف قابل ملاحظه باشد یعنی یک روش اتفاق افتاده است. درواقع، با تبدیل نگاره به جریان، مسئله شناسایی روش مفهومی به پیدا کردن داده‌های پرت در جریان تبدیل شده است. با توجه به اینکه برای ساختن ماتریس روابط، می‌توان از انواع معیارهای وابستگی موجود در فرایند کاوی استفاده کرد، معیار ^۲ قابل تعمیم است [14]. مشکل اصلی این روش، همچون روش‌های پیشین، وابستگی آن بهاندازه پنجره است. روش مطرح شده در [15] بدین صورت است که دو پنجره متولی تطبیقی را بر روی زیرنگاره‌ها حرکت می‌دهد. در هر گام، الگوریتم کشف فرایند ابتكاری^۳ را اعمال کرده و مدل فرایند حاصل از دو پنجره را استخراج می‌کند. بهمنظور شناسایی روش‌های ناگهانی، معیارهای گراف را از پنجره‌ها استخراج کرده و آزمون آماری را بر روی آن‌ها اعمال می‌کند. اگر مدل فرایند حاصل از پنجره نخست با مدل فرایند حاصل از پنجره دوم خیلی متفاوت باشد، به این معناست که تغییر رخ داده

۳- روش پیشنهادی

در این بخش بعد، به منظور برطرف کردن چالش‌های موجود، یک روش جدید مستقل از پنجره ارائه خواهیم کرد. در این روش، با مطرح کردن ایده تعییه دنباله سعی در شناسایی روش‌های ناگهانی در دیدگاه کنترل جریان داریم. با استفاده از تعییه دنباله، می‌توانیم روابط میان دنباله‌ها را به طور خودکار استخراج کنیم. روش پیشنهادی، دقت روش‌های موجود را بهبود خواهد داد.

^۳ Word embedding

^۱ Histogram

^۲ Heuristic miner

میانگین‌گیری شده و به لایه تجسم^۴ نگاشت داده می‌شوند؛ سپس با استفاده از وزن‌های ماتریس وزن خروجی، برای هر واژه یک امتیاز که نشان‌دهنده احتمال هدف‌بودن آن واژه است، محاسبه می‌شود. یک توالی از واژگان تعلیم W_1, W_2, \dots, W_T و پنجره C را در نظر بگیرید. هدف مدل CBOW بیشینه‌کردن احتمال میانگین لگاریتم رابطه زیر است [24]:

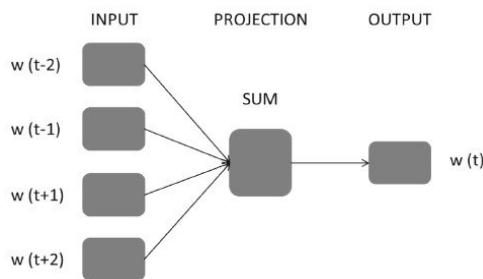
$$\frac{1}{T} \sum_{t=1}^T \log p(W_t | W_{t-c} \dots W_{t+c}) \quad (1)$$

که احتمال $p(W_t | W_{t-c} \dots W_{t+c})$ با استفاده ازتابع softmax (یک مدل طبقه‌بندی log-linear) محاسبه می‌شود:

$$p(W_t | W_{t-c} \dots W_{t+c}) = \frac{\exp(\bar{v}_w^T v'_w)}{\sum_{w=1}^V \exp(\bar{v}_w^T v'_w)} \quad (2)$$

که \bar{v}_w بردار خروجی واژه w ، V فرهنگ لغات واژگان و \bar{v} بردار ورودی میانگین گرفته‌شده از همه واژگان متن است [24]:

$$\bar{v} = \frac{1}{2c} \sum_{-c \leq j \leq c, j \neq 0} v_{W_{t+j}} \quad (3)$$



(شکل-۲): معماری مدل CBOW
(Figure-2): Architecture of the CBOW model

• مدل Skip-gram

این مدل برخلاف مدل CBOW سعی در پیش‌بینی واژگان متن با استفاده از واژگان هدف دارد. معماری آن در شکل (۳) نشان داده شده است. یک توالی از واژگان تعلیم W_1, W_2, \dots, W_T و پنجره C را در نظر بگیرید. هدف مدل Skip-gram بیشینه‌کردن احتمال میانگین لگاریتم رابطه زیر است [24]:

⁴ Projection

پنجره (تطبیقی یا ثابت) همه انواع روابط میان دنباله‌ها را به صورت خودکار استخراج کرده و دقت روش را بهمنظور شناسایی رانش فرایند در مقایسه با روش‌های موجود افزایش دهد.

در ادامه ابتدا مفاهیم اصلی در حوزه تعبیه واژه مطرح می‌شود و سپس ساختار روش پیشنهادی و اجزای آن با جزئیات گفته خواهد شد.

۱-۳-۱- تعبیه واژه

در تعبیه واژه، تمام واژگان موجود در فرهنگ لغت به بردار مدل می‌شوند. دو نوع تعبیه واژه وجود دارد [16]. ۱) تعبیه مبتنی بر تکرار [17]. ۲) تعبیه مبتنی بر پیش‌بینی [18]. [19]. بردارهای مبتنی بر تکرار، بر اساس هم‌خدادی یک واژه با گروهی خاصی از واژگان دیگر ساخته می‌شود؛ بنابراین، بردارهای تولیدشده پراکندگی^۱ زیاد و همچنین ابعاد بالایی دارند [20]؛ اما در روش‌های مبتنی بر پیش‌بینی، بردارهای تولیدشده متراکم^۲ و با ابعاد کم هستند. هدف اصلی روش‌های مبتنی بر پیش‌بینی این است که احتمال ظاهرشدن توالی خاصی از واژگان در یک متن را تخمین بزنند. در این مدل فرض شده است که واژگان نزدیک در یک توالی، از نظر آماری وابستگی بیشتری نسبت به یکدیگر دارند. یکی از مشهورترین روش‌های مبتنی بر پیش‌بینی، تعبیه واژه است [21]. تعبیه واژه یک شبکه عصبی دولایه است که متن را پردازش می‌کند. ورودی آن یک پیکره خام و خروجی آن یک فرهنگ لغت است که هر واژه به یک بردار مدل شده است. بردارهای واژگان مشابه به فضای برداری مشابهی نگاشت داده می‌شوند [22]. به عنوان مثال، واژه تهران و رم در فضای برداری نزدیک به هم قرار می‌گیرند؛ در حالی که بردار دو واژه تهران و سیب از هم دور است. دو الگوریتم اصلی برای تعبیه واژه وجود دارد: کیف لغات پیوسته^۳ و [23] Skip-gram.

• کیف لغات پیوسته

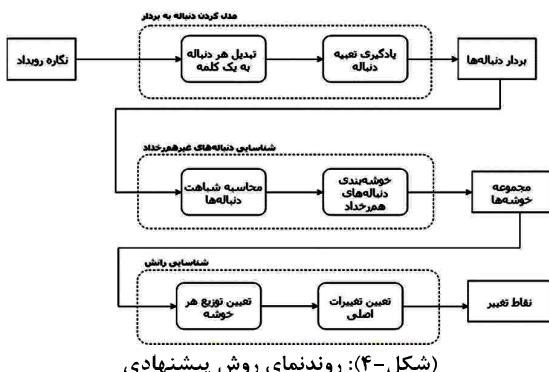
مدل CBOW با استفاده از واژگان متن موجود در یک پنجره معین، واژگان هدف را پیش‌بینی می‌کند. معماری این مدل در شکل (۲) نشان داده شده است. لایه ورودی، از واژگان متن موجود در پنجره تشکیل شده است که در آن بردارهای ورودی از ماتریس وزن ورودی بازیابی و

¹ Sparse

² Dense

³ Continuous bag-of-word

طرح می‌کنیم که قادر است به طور خودکار تمام انواع روابط میان دنباله‌ها را استخراج نماید.



(شکل-۴): روند نمایی روش پیشنهادی
(Figure-4): Architecture of the proposed approach

الف) تبدیل هر دنباله به یک واژه

با درنظر گرفتن فایل نگاره به عنوان متن و دنباله‌های فرایند به عنوان واژگان و با استفاده از روش تعبیه واژه، می‌توان ایده تعبیه دنباله را تفسیر کرد. هدف از ایده تعبیه دنباله، تولید بردار برای دنباله‌های موجود در نگاره رویداد است؛ به طوری که روابط میان برداها در فضای برداری نشان‌دهنده روابط میان دنباله‌ها در نگاره باشد.

به منظور به کارگیری ایده تعبیه دنباله، ابتدا باید فایل رویداد را آماده کرد؛ از این‌رو، هر دنباله در نگاره را با حذف فاصله میان فعالیت‌هایش به یک واژه تبدیل می‌کنیم. به عنوان مثال، دنباله مربوط به نمونه ۱ در شکل (۵) به صورت [abcdefg] تبدیل می‌شود؛ سپس جریانی از دنباله‌ها تولید می‌کنیم که در آن دنباله‌ها بر اساس برچسب زمانی نخستین رویدادشان مرتب شده‌اند.

ب) یادگیری تعبیه دنباله

تعبیه دنباله با یادگیری^۱ جریان تولید شده، برای هر دنباله (یک واژه‌ای) یک بردار از مقادیر عددی در فضای ویژگی تولید می‌کند. درنتیجه، با کمک بردارهای تولید شده، می‌توان جریان دنباله را به جریانی از بردارها با ابعاد یکسان تبدیل کرد که بیان‌گر اطلاعاتی درباره آمارهای هم‌رخدادی دنباله‌ها است.

شناختی دنباله‌های غیرهم‌رخداد

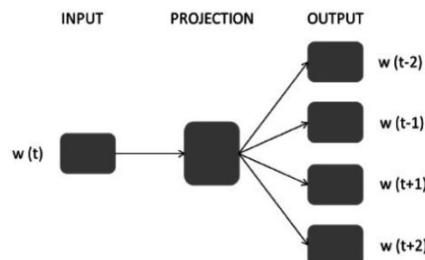
همان‌طور که پیش‌تر گفته شد، روش پیشنهادی یک روش مستقل از پنجره و مستقل از نوع فرایند ورودی است که می‌تواند به صورت خودکار روابط میان دنباله‌ها و رویدادها را استخراج کند و دقت شناسایی رانش فرایند در نگاره‌های کسب‌وکار را افزایش دهد. شکل (۴) روند نمایی روش پیشنهادی را نشان می‌دهد. روش پیشنهادی از سه گام اصلی مدل‌کردن دنباله‌ها به بردار، شناختی دنباله‌های غیرهم‌رخداد و شناختی رانش تشکیل شده است که در ادامه هر گام با جزئیات بیان می‌شود.

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(W_{t-j} | W_t) \quad (4)$$

که احتمال $p(W_{t+j} | W_t)$ با استفاده ازتابع softmax محاسبه می‌شود:

$$p(W_o | W_i) = \frac{\exp(\bar{v}_{W_o}^T v'_{W_i})}{\sum_{w=1}^V \exp(v_w^T v_{W_i})} \quad (5)$$

که v_w و v'_{W_o} بردارهای ورودی و خروجی واژه W فرهنگ لغت از واژگان است.



(شکل-۳): معماری مدل Skip-gram
(Figure-3): Architecture of the Skip-gram model

۲-۳-۲- ساختار روش پیشنهادی

همان‌طور که پیش‌تر گفته شد، روش پیشنهادی یک روش مستقل از پنجره و مستقل از نوع فرایند ورودی است که می‌تواند به صورت خودکار روابط میان دنباله‌ها و رویدادها را استخراج کند و دقت شناسایی رانش فرایند در نگاره‌های کسب‌وکار را افزایش دهد. روش پیشنهادی از سه گام اصلی مدل‌کردن دنباله‌ها به بردار، شناختی دنباله‌های غیرهم‌رخداد و شناختی رانش تشکیل شده است که در ادامه هر گام با جزئیات بیان می‌شود.

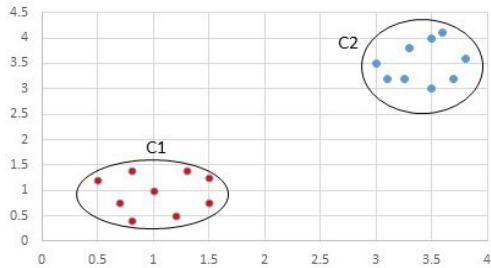
۲-۳-۲-۱- مدل کردن دنباله‌ها به بردار

بیش‌تر روش‌های موجود به منظور شناسایی رفتار فرایند و مشخص کردن روابط بین فعالیت‌ها، نیاز به تعریف مجموعه‌ای از ویژگی‌ها دارد. تعداد و نوع ویژگی‌های انتخاب شده در دقت روش‌ها تأثیرگذار و حتی در بعضی از موارد نیاز است که کاربر از یک دانش قبلی نسبت به فرایند در حال اجرا بهره‌مند باشد تا ویژگی مناسبی را انتخاب کند. به منظور حذف این محدودیت، ما مفهوم تعبیه دنباله را

¹ Train



همان‌طور که در شکل (۶) نشان داده است، همه اعضای خوشه C1 در فضای ویژگی یکسان و در عین حال متفاوت با اعضای خوشه C2 قرار دارند.



(شکل-۶): تصویرسازی دوبعدی از فضای ویژگی. با توجه به این

شکل اعضای هر خوشه در فضای ویژگی مشابه قرار دارند.

(Figure-6): Two-dimensional visualization of feature space. The figure illustrates that the members of each cluster are close to each other.

شناسایی رانش

هدف از این مرحله، استفاده از خوشه‌های تولید شده به منظور شناسایی رانش فرایند و تعیین مکان دقیق وقوع آن است. در ادامه به منظور شناسایی رانش‌ها با بیشترین دقت ممکن، روش‌های خاصی اعمال شده است.

الف) تعیین توزیع هر خوشه

ابتدا برای هر خوشه یک بردار توزیع تولید می‌کنیم که نشان‌دهنده نحوه توزیع دنباله‌های هر خوشه در نگاره رویداد است. طول بردارهای توزیع، برابر با تعداد نمونه‌های نگاره است که یک به معنای حضور و صفر به معنای عدم حضور اعضای یک خوشه در بردار توزیع آن است. شکل ۷ نحوه توزیع دو خوشه C1 و C2 را نشان می‌دهد که نواحی مترافق متناظر دنباله‌های هر خوشه در نگاره از هم جدا شده اند. طول بردارهای توزیع در شکل (۷) برابر با ۲۵۰۰ است یا به عبارت دیگر نگاره دارای ۲۵۰۰ نمونه است.

ب) تعیین تغییرات اصلی

به منظور تعیین مکان‌های پرتکرار، بردارهای توزیع را به فضای فوریه منتقل می‌کنیم. در حقیقت، با به کارگیری تبدیل فوریه می‌توانیم یک سیگنال را به فرکانس‌های مختلف آن تجزیه کنیم [25].

با توجه به این که فرکانس‌های پایین نشان‌دهنده تغییرات اصلی در یک سیگنال و فرکانس‌های بالا به طور عمومی نویه هستند، پس با اعمال یک فیلتر پایین‌گذر^۱ می‌توان تغییرات اصلی را شناسایی و نویه‌ها را حذف کرد. در اینجا منظور از نویه، دنباله‌هایی هستند که به اشتباه در یک خوشه قرار گرفته‌اند.

^۱ Low pass filter

Case id	Activity	Time stamp
1	a	2004/02/11 13:00:00.000
	b	2004/02/11 13:25:28.91
	c	2004/02/11 13:28:18.693
	d	2004/02/11 13:42:06.935
	e	2004/02/11 14:52:58.531
	f	2004/02/11 15:00:07.055
	g	2004/02/11 15:06:39
	h	2004/02/11 15:06:39.125
.	.	.
n	.	.
	a	2004/03/02 17:30:00.000
	b	2004/03/02 18:00:06.067
	r	2004/03/02 18:03:27.651
	s	2004/03/02 18:18:29.374
	b	2004/03/02 18:49:11.161
	c	2004/03/02 19:31:05.327
	d	2004/03/02 19:38:43.799
	e	2004/03/02 19:44:37.811
	f	2004/03/02 19:59:02.032
	g	2004/03/03 13:37:11.601
	h	2004/03/03 13:37:11.601

(شکل-۵): قسمتی از یک نگاره

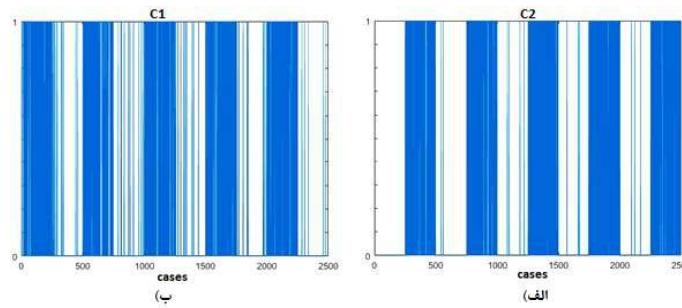
(Figure-5): A fragment of an event log

الف) محاسبه شباهت دنباله

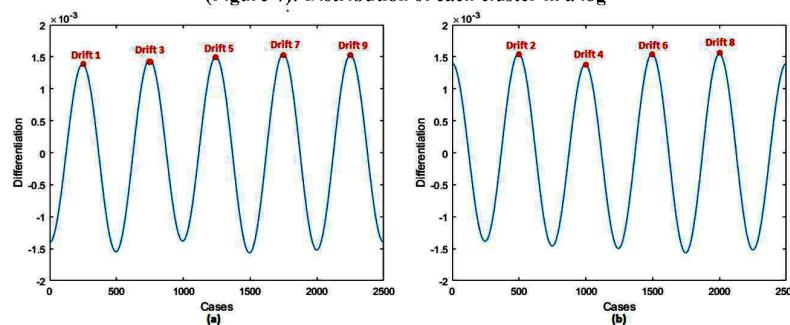
با محاسبه شباهت برداری میان دنباله‌های نگاره و شناسایی دنباله‌ها با کمترین شباهت یا بیشترین فاصله می‌توان دنباله‌های غیر هم‌رخداد را مشخص کرد. به عبارت دیگر، هر تغییر رفتار فرایند از A به B منجر به افزایش فاصله یا کاهش هم‌رخدادی در میان مجموعه نمونه‌های فرایند A و مجموعه نمونه‌های فرایند B می‌شود و درنتیجه بردار دنباله‌ها قبل از نقطه تغییر (وقوع رانش) با بردار دنباله‌ها بعد از آن نقطه متفاوت است. به منظور شناسایی مجموعه دنباله‌های غیر هم‌رخداد، از شباهت کسینوسی میان بردارها استفاده شده است.

ب) خوشبندی دنباله‌های هم‌رخداد

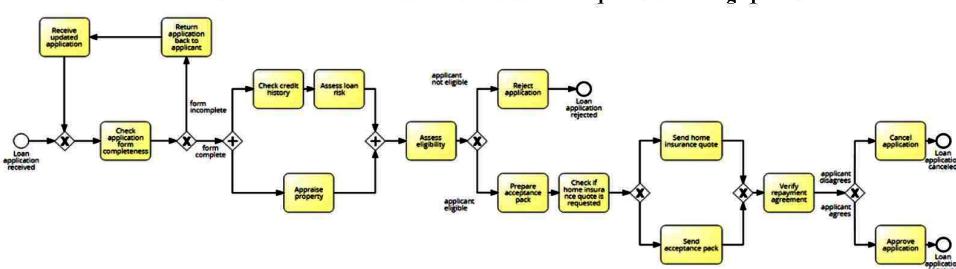
در مرحله بعد، نیاز داریم از میان مجموعه دنباله‌های غیر هم‌رخداد شناسایی شده، مجموعه‌ای هم‌رخداد را شناسایی کنیم. علت این امر که چرا از ابتدا مجموعه دنباله‌های هم‌رخداد شناسایی نمی‌شوند این است که ابتدا باید دنباله‌هایی را که در موقع رانش تأثیرگذار نیستند، حذف کنیم؛ بنابراین، پس از شناسایی مجموعه‌های غیر هم‌رخداد، به منظور شناسایی مجموعه دنباله‌ها با بیشترین هم‌رخدادی از خوشبندی سلسله‌مراتبی استفاده می‌کنیم. در ابتدا هر دنباله در یک خوشه قرار دارد و در هر مرحله دو خوشه با بیشترین هم‌رخدادی ادغام می‌شوند؛ درنهایت، خروجی این مرحله مجموعه‌ای از خوشه‌ها است که اعضای هر خوشه در فضای برداری مشابه و در مقابل، در فضای برداری نامشابه با تمام اعضای خوشه‌های دیگر قرار دارند؛ همچنین، هر دنباله تنها می‌تواند در یک خوشه قرار بگیرد و اشتراک خوشه‌ها تهی است. اگر فرض کنیم که بردارها دوبعدی هستند و خروجی خوشبندی سلسله‌مراتبی دو خوشه C1 و C2 است،



(شکل-۷): نحوه توزیع هر خوشه در نگاره (Figure-7): Distribution of each cluster in a log



(شکل-۸): (الف) نرخ تغییرات در بردار توزیع خوشه C1. (ب) نرخ تغییرات در بردار توزیع خوشه C2. قله‌ها نشان‌دهنده تغییر هستند. (Figure-8): (a) Rate of changes in the distribution vector of cluster C1. (b) Rate of changes in the distribution vector of cluster C2. Remarkable peaks are change points.



(شکل-۹): مدل BPMN مربوط به فرایند وام (Figure-9): Base BPMN model of loan application process

۴- ارزیابی نتایج

به منظور ارزیابی روش پیشنهادی از مجموعه داده مطرح شده در [12] استفاده شده است.^۱ [12] ۷۲ نگاره با ساختار متفاوت ارائه داده است. شکل (۹) مدل فرایند پایه در همه نگاره‌ها را نشان می‌دهد. این مدل مربوط به فرایند درخواست وام است که شامل پاتریه فعالیت، یک رویداد آغازین و سه رویداد خاتمه است.

به منظور ایجاد رانش در این نگاره‌ها، دوازده گروه تغییر متفاوت به مدل فرایند پایه اعمال شده است. گروه‌ای تغییر در جدول (۱) نشان داده شده است. این گروه‌ها، بیان‌کننده تغییرات ممکن بر مدل فرایند کسب‌وکار هستند، اضافه یا حذف یا حلقوی کردن یک قطعه مدل،

پس از اعمال فیلتر پایین‌گذر، بردارهای توزیع تغییریافته را به فضای زمان برگردانده و برای تشخیص نرخ تغییرات از آن مشتق می‌گیریم. نقطه‌ها با بیشترین نرخ تغییر یا به عبارت دیگر قله‌ها نشان‌دهنده مکان وقوع رانش فرایند در نگاره هستند. شکل (۸) خروجی مشتق گرفته شده از بردارهای توزیع شکل (۷) را نشان می‌دهد که نقاط قله نشان‌دهنده مکان وقوع رانش هستند. بنابراین، نقاط ۲۵۰، ۲۲۵۰، ۲۰۰۰، ۱۷۵۰، ۱۵۰۰، ۱۲۵۰، ۱۰۰۰، ۷۵۰، ۵۰۰ به عنوان شاخص دنباله‌هایی که در آن‌ها رانش رخ داده است شناسایی می‌شوند؛ بنابراین با ارائه روشی جدید در شناسایی رانش مفهومی، توانستیم با حل مشکلات پیشین، تغییرات را شناسایی کنیم.

^۱ <http://apromore.org/platform/tools>

مواجهه با هر نگاره به صورت جدگانه با روش‌های دیگر مقایسه می‌شود. دو معیار امتیاز- F و متوسط تأخیر به عنوان معیارهای ارزیابی دقت در روش‌های شناسایی رانش مفهومی در حوزه داده‌کاوی ارائه شده‌اند [26]. این معیارها برای ارزیابی روش‌های موجود در حوزه فرایند کاوی نیز استفاده شده و لذا امکان مقایسه روش پیشنهادی را فراهم می‌آورد. معیار امتیاز- F ، متوسط هارمونیک دقت^۱ و فراخوانی^۲ را اندازه می‌گیرد که به صورت زیر محاسبه می‌شوند.

$$precision = \frac{TP}{TP + FP} \quad (6)$$

$$recall = \frac{TP}{TP + FN} \quad (7)$$

$$F-score = \frac{2 \times precision \times recall}{precision + recall} \quad (8)$$

منظور از مثبت صحیح^۳، منفی کاذب^۴ و مثبت کاذب^۵ در هر حوزه متفاوت است. در زمینه شناسایی رانش فرایند این مفاهیم به صورت زیر تعریف می‌شوند:

- مثبت صحیح (TP): تعداد نقاطی که به درستی به عنوان رانش شناسایی شده‌اند.

- منفی کاذب (FN): تعداد نقاطی که به عنوان رانش شناسایی نشده‌اند، اما در واقعیت وجود دارند.

- مثبت کاذب (FP): تعداد نقاطی که به عنوان رانش شناسایی شده‌اند، اما در واقعیت وجود ندارند.

متوسط تأخیر معیار دیگری برای ارزیابی دقت است که بیان کننده متوسط تعداد دنباله‌ها میان نقطه واقعی تغییر و نقطه شناسایی شده است. هر چه متوسط تأخیر یک روش کمتر باشد، عملکرد آن در شناسایی رانش فرایند بهتر خواهد بود.

در این پژوهش از روش ارزیابی مقایسه‌ای استفاده شده است. از میان روش‌های یادشده در بخش کارهای گذشته، روش‌های مطرح شده در [3, 15, 12] جهت مقایسه و ارزیابی دقت روش پیشنهادی استفاده شده‌اند. با توجه به این‌که ارزیابی انجام‌شده در این روش‌ها، بر روی مجموعه‌داده یکسانی است، نیاز به پیاده‌سازی مجدد آن‌ها نبود. همان‌طور که پیش‌تر گفته شد، این مجموعه‌داده یکسانی، مجموعه‌داده ارائه شده در [12] است.

¹ Precision

² Recall

³ True positive

⁴ False negative

⁵ False positive

جایه‌جاکردن دو قطعه مدل، یا موازی‌سازی دو قطعه متولی. [12] الگوهای تغییر را در سه دسته افزودن (I)، تغییر ترتیب (R) و اختیاری (O) تقسیم‌بندی کرده و با ترکیب کردن آن‌ها شش الگو تغییر پیچیده‌تر را تولید می‌نماید. به عنوان مثال، الگوی ترکیبی IOR را می‌توان با اضافه کردن یک فعالیت جدید (I)، موازی قراردادن این فعالیت با یک فعالیت موجود (O) و درنهایت قراردادن آن در ساختار حلقوی (R) تولید کرد.

(جدول-۱): الگوهای تغییر در دیدگاه کنترل جریان

(Table-1): Simple control-flow change patterns

کد	الگو تغییر	دسته
re	حذف یا اضافه کردن قطعه	I
cf	قراردادن دو قطعه به صورت شرطی / متولی	R
lp	قراردادن یا خارج کردن قطعه در / از حلقه	O
p1	قراردادن دو قطعه به صورت متولی / موازی	R
cb	اضافه کردن یا حذف قابلیت پرش از قطعه	O
cm	قراردادن یا خارج کردن قطعه در / از شاخه شرطی	I
cd	همگام‌سازی دو قطعه	R
cp	تکرار قطعه	I
pm	قرارداد یا خارج کردن قطعه در / از شاخه موازی	I
rp	جایگزینی قطعه	I
sw	جایه‌جاکردن دو قطعه	I
fr	تغییر تکرار شاخه‌ها	O

[12] به منظور قرار دادن رانش‌ها در فواصل متفاوت، ۴ نگاره با ۲۵۰۰، ۵۰۰۰، ۷۵۰۰ و ۱۰۰۰۰ دنباله برای مدل فرایند پایه و برای هر یک از هجده الگو تغییر تولید کرده که در مجموع برابر با ۷۲ نگاره خواهد شد. در هر نگاره ۹ رانش، در هر ده درصد از اندازه آن، قرار داده شده است. با آگاهی از تعداد و مکان هر رانش در نگاره‌ها، می‌توان یک استاندارد طلایی تولید و دقت روش را با آن ارزیابی کرد. گفتنی است که هدف از ایجاد این نگاره‌های ساختگی تولید مجموعه‌ای از داده‌های آموزش و آزمون نیست؛ بلکه به منظور ارزیابی دقت روش ارائه شده لازم است نگاره‌هایی تولید شود که در بردارنده انواع تغییرات ممکن باشند. از این‌رو، هر نگاره به صورت جدگانه و به صورت تدریجی آموزش داده و دقت عملکرد روش پیشنهادی در



معیارهای امتیاز- F و متوسط تأخیر برای هر یک از الگوهای تغییر محاسبه شده است. نتایج دقیق این معیارها برای هر الگو در روش پیشنهادی بهدست آمده و بهمراه نتایج روش‌های دیگر در جدول (۲) نشان داده شده است. نتایج این جدول به ازای هر الگوی تغییر بر روی چهار نگاره متفاوت میانگین گرفته شده است. نمودارهای شکل‌های (۱۰ و ۱۱) نیز بهترتب مقایسه روش پیشنهادی با روش‌های پایه را از نظر امتیاز- F و متوسط تأخیر بهازی هر الگو تغییر (میانگین گرفته شده بر روی چهار نگاره) نشان می‌دهند. همان‌طور که در نمودار شکل (۱۰) مشاهده می‌کنید، روش پیشنهادی به امتیاز- F برابر یک برای همه الگوها بهجز OTR دست یافته است که در مقایسه با روش‌های پایه بهبود فراوانی را نشان می‌دهد. همچنین، روش پیشنهادی در سیزده الگو تغییر به تأخیر کمتر از بیست دنباله، در سه الگو تغییر به تأخیر کمتر از سی دنباله و تنها در دو الگو تغییر به تأخیر بیشتر از سی دنباله دست یافته که در مقایسه با روش‌های دیگر، در شانزده الگو، متوسط تأخیر را در شناسایی مکان رانش بهطور مناسبی

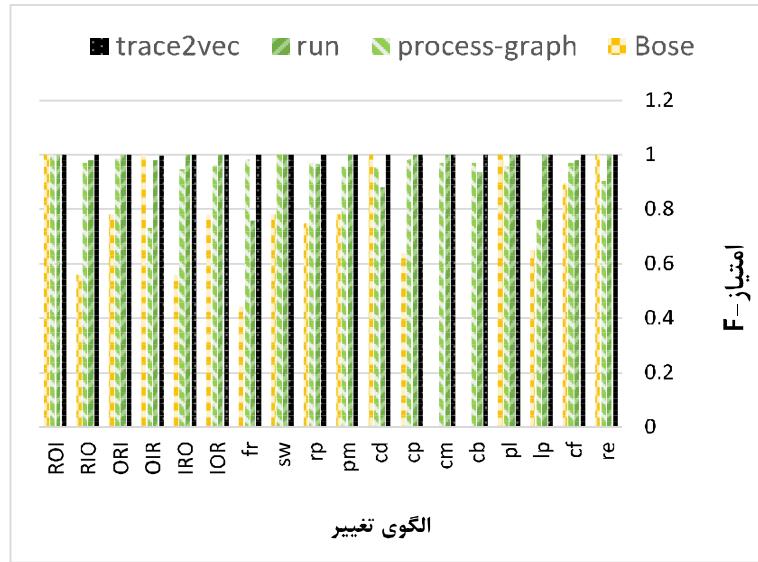
کاهش داده است. به طور میانگین، امتیاز- F تقریبی بهدست آمده برای ۷۲ نگاره مورد بررسی برابر با ۰.۹۹، ۰.۹۷ و ۰.۹۰ بهترتب برای روش پیشنهادی، run و process-graph است که روش پیشنهادی بهبود قابل توجهی نسبت به روش‌های دیگر دارد؛ همچنین، متوسط تأخیر تقریبی بهدست آمده برای ۷۲ نگاره مورد بررسی، به طور میانگین برابر با ۱۳، ۲۲ و ۴۷ دنباله بهترتب برای روش پیشنهادی، process-graph و run است. بعبارت دیگر، به طور میانگین، مکان رانش شناسایی شده توسط روش پیشنهادی حدود سیزده دنباله با مکان واقعی رانش فاصله دارد که سه روش دیگر را بهبود داده است.

در آزمایش بعدی، روش پیشنهادی را بر روی نگاره دنیای واقعی اعمال کردیم. نگاره مورد استفاده مربوط به پنج منطقه مختلف از شهرداری مشهد است. در ابتدا، به منظور بالابردن کیفیت تحلیل، پیش‌برداش‌هایی برای مدیریت نویسه‌ها، مقادیر ازدست‌رفته و همچنین ناسازگاری‌ها بر روی پایگاه داده انجام شد.

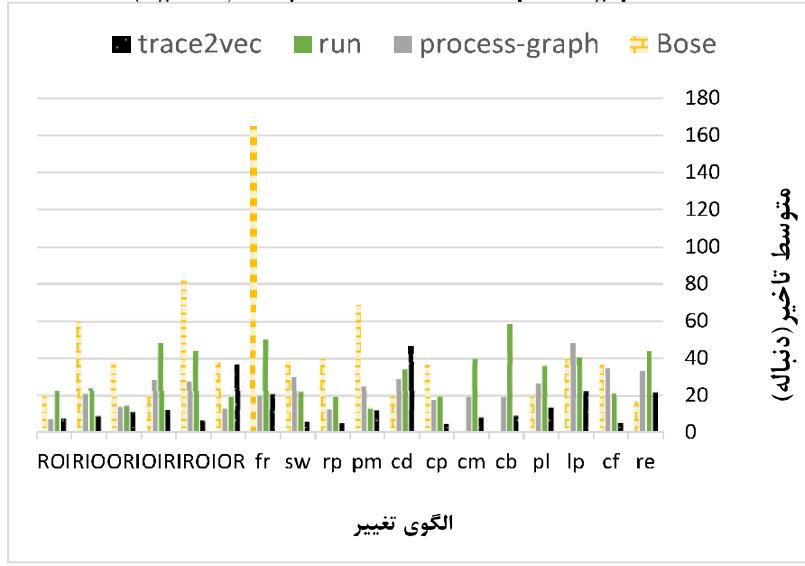
(جدول-۲): مقادیر دقیق روش پیشنهادی از دو دیدگاه امتیاز- F و متوسط تأخیر در مقایسه با روش‌های run و process-graph و Bose

(Table-2): Comparison of average F-score and mean delay values

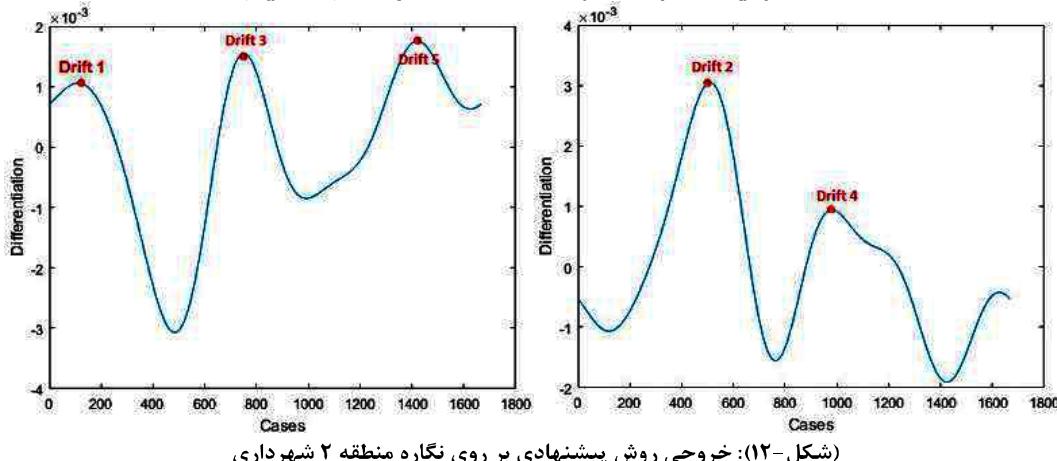
الگو تغییر	trace2vec		run [۱۲]		process graph [۱۵]		Bose [۳]	
	امتیاز- F	متوسط تأخیر	امتیاز- F	متوسط تأخیر	امتیاز- F	متوسط تأخیر	امتیاز- F	متوسط تأخیر
Re	1	21.83	1	44.03	0.9036	33.02	1	17
Cf	1	5.08	0.9824	21	0.9853	34.62	0.8950	36
Lp	1	21.69	1	40.29	0.7618	48.03	0.6484	41
Pl	1	13.8	1	35.74	0.9575	26.33	1	20
Cb	1	9.1	0.9387	58.55	0.9722	18.94	0	0
Cm	1	8.52	1	39.85	0.9722	19.24	0	0
Cp	1	4.58	1	19.66	0.9853	17.59	0.6394	36
Cd	1	46.44	0.8799	34.62	0.9546	28.62	1	20
Pm	1	11.97	1	12.88	0.9869	24.78	0.7804	69
Rp	1	4.86	0.9666	19.18	0.9722	12.67	0.75	40
Sw	1	6.02	1	21.67	1	29.61	0.7804	39
Fr	1	20.5	0.7569	49.92	0.9853	19.92	0.4420	165
IOR	1	36.66	1	19.11	0.9606	13	0.7804	38
IRO	1	6.8	1	43.96	0.9487	27.22	0.5611	82
OIR	0.9967	12.11	0.9803	47.89	0.7331	28.06	1	20
ORI	1	11.38	1	14.51	0.9869	14.25	0.7804	38
RIO	1	9.08	0.9824	23.81	0.9722	20.77	0.5611	60
ROI	1	7.66	1	22.51	1	7.31	1	20
میانگین	0.9998	12.15	0.9715	31.62	0.9466	23.56	0.7010	46.31



(شکل-۱۰): امتیاز- F -به دست آمده از روش پیشنهادی در مقایسه با run، trace2vec و process-graph
 (Figure-10): Comparison of F-score values per change pattern



(شکل-۱۱): متوسط تأخیر به دست آمده از روش پیشنهادی در مقایسه با run، trace2vec و process-graph
 (Figure-11): Comparison of mean delay values per change pattern



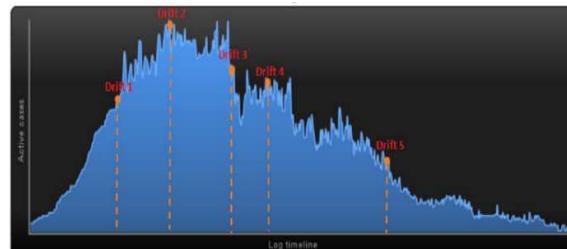
(شکل-۱۲): خروجی روش پیشنهادی بر روی نگاره منطقه ۲ شهرداری
 (Figure-12): Plots of the output results after applying the trace2vec approach to M2

- پیشنهادی بر روی نگاره واقعی نیز نشان داده است. با وجود پیشرفت‌های زیادی که در زمینه شناسایی رانش فرایند انجام شده است، اما هنوز موضوعات باز برای پژوهش در این راستا وجود دارد. برخی از آن‌ها عبارت‌اند از:
- **ارائه روشی بهمنظور شناسایی رانش‌ها در دیدگاه داده و منبع:** هیچ یک از روش‌های موجود توانایی شناسایی تغییرات در دیدگاه منبع و با دقت مناسب در دیدگاه داده را ندارند؛ ازین‌رو، ارائه روشی با کارایی مناسب، اطلاعات ارزشمندی در این دو دیدگاه ارائه خواهد کرد.
 - **ارائه روشی بهمنظور شناسایی رانش‌های تدریجی با دقت بالاتر:** همان‌طور که پیش‌تر گفته شد، تنها دو روش مطرح شده در [7] و [13] توانایی شناسایی رانش‌های تدریجی را دارند. با این حال، نتایج ارائه‌شده توسط آن‌ها دقت مناسب را نشان نمی‌دهد؛ بنابراین، روشی جدید بهمنظور افزایش دقت در شناسایی رانش‌های تدریجی حائز اهمیت است.
 - **ارائه روشی بهمنظور شناسایی رانش‌های متناوب و افزایشی:** شناسایی رانش‌های متناوب و افزایشی با دقت مناسب، اطلاعات ارزشمندی درباره عوامل تأثیرگذار بر فرایند سازمان ارائه می‌دهد.

6- References

- [1] W. Van Der Aalst, A. Adriansyah, A. K. A. Do Medeiros, F. Arcieri, T. Baier, T. Blickle, et al., "Process mining manifesto", in *International Conference on Business Process Management*, vol. 37 No.3, pp. 169-194, 2011.
- [2] M. U. Lavanya and M. S. K. Talluri, "Dealing With Concept Drifts In Process Mining Using Event Logs", *International Journal Of Engineering And Computer Science*, vol. 4, pp. 13433-13437, 2015.
- [3] R. J. C. Bosc, W. M. Van Der Aalst, I. Žliobaitė, and M. Pechenizkiy, "Dealing with concept drifts in process mining", *IEEE transactions on neural networks and learning systems*, vol. 25 No.1, pp. 154-171, 2014.
- [4] J. C. Schlimmer and R. H. Granger, "Beyond Incremental Processing: Tracking Concept Drift", *the Association for the Advancement of Artificial Intelligence*, pp. 502-507, 1986.
- [5] R. J. C. Bosc, W. M. van der Aalst, I. Žliobaitė, and M. Pechenizkiy, "Handling concept drift in process mining", *23rd International Conference on Advanced Information Systems*

به دلیل محدودیت فضای نتایج حاصل بر روی یک منطقه شهرداری ارائه شده است. شکل (۱۲) نتایج حاصل از اعمال روش پیشنهادی بر روی M2 را نشان می‌دهد. پنج رانش در اندیس‌های ۱۱۱، ۵۱۰، ۷۵۱، ۹۸۱ و ۱۴۲۱ برای شهرداری منطقه ۲ شناسایی شده، این درحالی است که با اعمال روش مطرح شده در [12] هیچ رانشی شناسایی نشد. نتایج به دست آمده را با سه تحلیل‌گر کسب‌وکار مورد ارزیابی قرار دادیم. نتیجه ارزیابی‌ها حاکی از عملکرد صحیح روش پیشنهادی بود. تحلیل‌گران کسب‌وکار علت اصلی این رانش‌ها را وجود دو تغییر در فرایند سازمان اعلام کردند: تغییرات مدیریتی و تغییرات نرم‌افزاری. آن‌ها علت اصلی تغییرات نرم‌افزاری را بهبود کارایی فرایند کسب‌وکار بیان کردند که منجر به کاهش تعداد نمونه‌ها در فرایند می‌شود. شکل (۱۳) تاثیر تغییرات شناسایی شده را بر روی توزیع تعداد نمونه‌ها در طی زمان نشان می‌دهد. همان‌طور که در شکل (۱۳) مشاهده می‌شود، اگرچه در برخی بازه‌ها تعداد نمونه‌ها افزایش یافته، اما درنهایت تعداد نمونه‌های متمایز در این منطقه از شهرداری کاهش یافته که این دلیلی بر بیبود کارایی فرایند است.



(شکل-۱۳): موقعیت رانش‌های شناسایی شده بر روی نمودار فعالیت‌ها در طی زمان برای نگاره منطقه ۲ شهرداری مشهد
(Figure-13): The position of the drifts detected by our approach shown in a plot of active cases vs log timeline for M2

۵- نتیجه‌گیری و کارهای آینده

در این مقاله، سعی شد با ارائه مفهومی جدید دقت روش‌های پیشین در شناسایی رانش فرایند را افزایش دهیم. با بیان ایده تعبیه دنباله و همچنین به کارگیری روش‌های خوشبندی و تبدیل فوریه روشی را ارائه کردیم که همه انواع روابط میان دنباله‌ها را به صورت خودکار استخراج و بدون استفاده از پنجره، رانش‌های ناگهانی را شناسایی کند. نتایج ارزیابی بر روی یک مجموعه داده استاندارد حاکی از این است که روش پیشنهادی منجر به افزایش دقت در شناسایی رانش فرایند شده است؛ همچنین، عملکرد روش

فصلنامه



- [16] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors", *ACL* (1), pp. 238-247, 2014.
- [17] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol.41, p. 391, 1990.
- [18] A. Mandelbaum and A. Shalev, "Word Embeddings and Their Use In Sentence Classification Tasks," arXiv preprint arXiv:1610.08229, 2016.
- [19] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations", *HLT-NAACL*, Atlanta, USA, pp.746-751, 2013.
- [20] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents", *ICML*, Beijing, China, pp. 1188-1196, 2014.
- [21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781, 2013.
- [22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", *the 27th Conference on Neural Information Processing Systems(NIPS)*, Nevada, United States, pp. 3111-3119, 2013.
- [23] X. Rong, "word2vec parameter learning explained", arXiv preprint arXiv:1411.273, 2014.
- [24] P. Ristoski and H. Paulheim, "Rdf2vec: Rdf graph embeddings for data mining", *15th International Semantic Web Conference(ISWC)*, Kobe, Japan, pp. 498-514, 2016.
- [25] M. Rahman, "Applications of Fourier transforms to generalized functions", WIT Press, 2011.
- [26] S.-S. Ho, "A martingale framework for concept change detection in time-varying data streams", *the 22nd international conference on Machine learning*, Germany, ACM, pp. 321-327, 2005.
- [27] H. Schonenberg, R. Mans, N. Russel, N. Mulyar, and W. van der Aalst, "Process flexibility: A survey of contemporary approaches", *4th International Workshop on Advances in Enterprise Engineering I*, Springer, pp. 16-30, 2008.
- [28] J. Martjushev, R. J. C. Bose, and W. M. van der Aalst, "Change Point Detection and Dealing with Gradual and Multi-order Dynamics in Process Mining", *14th International Conference on Business Informatics Research*, Tartu, Estonia, Springer, pp. 161-178, 2015.
- [29] R. Accorsi and T. Stocker, "Discovering workflow changes with time-based trace clustering", *International Symposium on Data-Driven Process Discovery and Analysis*, Italy, Springer, pp. 154-168, 2011.
- [30] B. Hompes, J. Buijs, W. van der Aalst, P. Dixit, and J. Buurman, 2015, "Detecting Change in Processes Using Comparative Trace Clustering", *the 5th International Symposium on Data-driven Process Discovery and Analysis (SIMPDA)*, Vienna, Austria , pp. 95-108.
- [31] J. Carmona and R. Gavalda, "Online techniques for dealing with concept drift in process mining", *11th International Symposium on Intelligent Data Analysis*, Finland, Springer, pp.90-102, 2012.
- [32] P. Weber, B. Bordbar, and P. Tiño, "Real-Time Detection of Process Change using Process Mining", *ICCSW*, United Kingdom, 2011.
- [33] A. Maaradji, M. Dumas, M. La Rosa, and A. Ostovar, "Fast and accurate business process drift detection", *the 13th International Conference on Business Process Management*, Austria, Springer, pp. 406-422, 2015.
- [34] A. Maaradji, M. Dumas, M. L. Rosa, and A. Ostovar, "Detecting Sudden and Gradual Drifts in Business Processes from Execution Traces", *IEEE Trans. Knowl. Data Eng.*, vol. 29, pp. 2140-2154, 2017.
- [35] T. Li, T. He, Z. Wang, Y. Zhang, and D. Chu, "Unraveling Process Evolution by Handling Concept Drifts in Process Mining", *in Services Computing (SCC)*, IEEE International Conference on, pp. 442-449, 2017.
- [36] A. Seeliger, T. Nolle, and M. Mühlhäuser, "Detecting Concept Drift in Processes using Graph Metrics on Process Graphs", *Proceedings of the 9th Conference on Subject-oriented Business Process Management*, 2017.

فاطمه خجسته دانش آموخته کارشناسی
ارشد مهندسی رایانه نرم افزار از دانشگاه
فردوسی مشهد و حوزه پژوهشی ایشان
فرآیند کاوی و داده کاوی است.

نشانی رایانمای ایشان عبارت است از:

fatemeh.khojasteh@mail.um.ac.ir



محسن کاهانی استاد تمام گروه رایانه و سرپرست آزمایشگاه فناوری وب در دانشگاه فردوسی مشهد است. حوزه‌های فعالیت ایشان وب‌معنایی، پردازش زبان طبیعی، فرایندکاوی، پردازش متن، تحلیل نظرات، داده‌های حجمی است. نشانی رایانمۀ ایشان عبارت است از:

kahani@um.ac.ir



بهشید بهكمال استادیار دانشگاه فردوسی مشهد و حوزه‌های فعالیت ایشان داده‌کاوی، فرایندکاوی و داده‌های پیوندی است. نشانی رایانمۀ ایشان عبارت است از:

behkamal@um.ac.ir