ارائه یک رویکرد فضایی برای انتزاع مدل فرآیند دریافت و تحویل بسته

یاسر صالح‌آبادی[1]، بهشید بهکمال[2]، مصطفی میرزایی[3]

[1] دانشجوی ارشد نرم‌افزار، گروه کامپیوتر دانشکده مهندسی، دانشگاه فردوسی مشهد
ysalehabadi@mail.um.ac.ir

[2] استادیار، گروه کامپیوتر دانشکده مهندسی، دانشگاه فردوسی مشهد
behkamal@um.ac.ir

[3] دانشجوی دکتری نرم‌افزار، گروه کامپیوتر دانشکده مهندسی، دانشگاه فردوسی مشهد
mostafa.mirzaie@mail.um.ac.ir

چکیده

تحلیل دقیق فرآیندهای کسب‌وکار همواره دغدغه مدیران و تحلیل‌گران سازمان‌ها بوده است. لازمه‌ی یک تحلیل خوب، استخراج مدل‌هایی قابل فهم و قابل درک از نگاره‌رویداد است. در دنیای واقعی، رویدادها غالباً با جزئیات زیاد و سطح ریزدانگی بالا توسط سیستم‌های اطلاعاتی ذخیره می‌شوند و اگر الگوریتم‌های کشف را بر روی چنین نگاره‌رویدادهایی اعمال کنیم، نتیجه مدل‌هایی غیرقابل درک و پیچیده خواهد بود. از این رو تکنیک‌های متنوعی تاکنون برای انتزاع رویدادها معرفی شده است. در دسته‌ای از فرآیندها، لازم است تا نحوه‌ی انجام فرآیند از جنبه‌ی تغییرات موقعیت مکانی مورد تحلیل قرار گیرد مثل فرآیندهای مربوط به دریافت و تحویل بسته. تکنیک ارائه شده در این مقاله با یک رویکرد فضایی و با دو شیوه‌ی مختلف به انتزاع رویدادها پرداخته است. در شیوه‌ی نخست، برچسب‌های جغرافیایی موجود در نگاره‌رویداد با استفاده از الگوریتم سلسله‌مراتبی خوشه‌بندی بر اساس نزدیکی فاصله گروه‌بندی شده و اعضای هر گروه در سطوح بالاتر تجمیع می‌شوند. در شیوه‌ی دوم، به منظور ایجاد سطوح دقیق‌تر و بامعناتری از انتزاع مثل انتزاع در سطح شهر و سطح استان، در حین خوشه‌بندی از فراداده مناسب استفاده می‌شود. نتایج آزمایشگاهی مربوط به این رویکرد نشان‌دهنده‌ی بهبود تحلیل مدل‌ها بر اساس معیار پیچیدگی است. همچنین رویکرد ارائه شده در این مقاله قادر است تا نقاطی را که بر اساس تقسیم‌بندی‌های دولتی مثل استان در یک خوشه قرار گرفته‌اند و بهتر است بر اساس معیار نزدیکی فاصله در خوشه‌ی دیگر قرار گیرند را شناسایی کند.

کلمات کلیدی

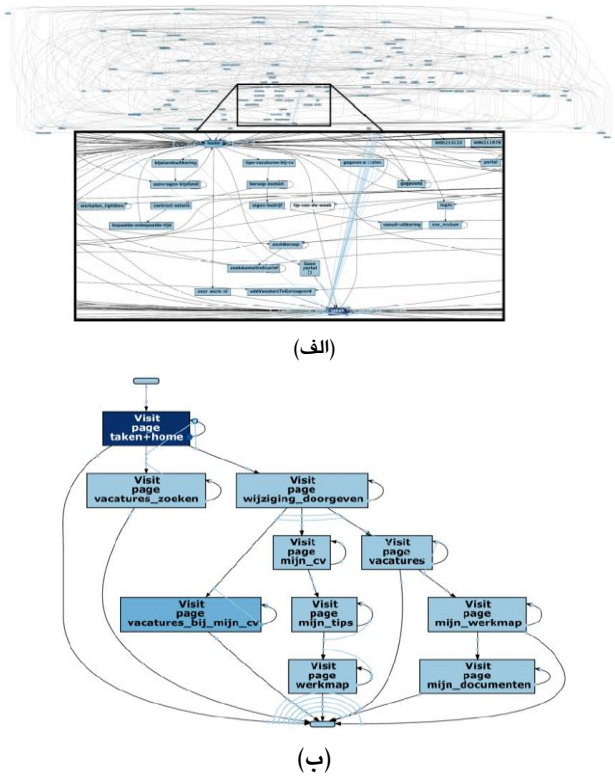
فرآیندکاوی، مدل‌سازی فرآیند، انتزاع رویداد، خوشه‌بندی سلسله‌مراتبی، موقعیت مکانی

۱- مقدمه

امروزه در بسیاری از فرآیندهای تجاری، سیستم‌های سازمانی، سیستم‌های اتوماسیون و کنترل، سیستم‌های پزشکی، فعالیت‌های روزانه، دستگاه‌های مرتبط به اینترنت اشیا و شبکه‌های اجتماعی رویدادها به همراه ویژگی‌هایشان ثبت می‌شوند. این رویدادها فرصت‌های بسیاری برای به دست آوردن دانش برای درک آنچه اتفاق می‌افتد ارائه می‌دهند. استفاده از این داده‌ها، به فرآیندکاوی منجر می‌شود که هدف از آن کشف، بررسی انطباق و بهبود فرآیندهای تجاری واقعی در بسیاری از سیستم‌ها است [1].

اجرای فرآیندها در سازمان‌ها، دنباله‌ای از داده‌های رویداد را ایجاد می‌کند که در سیستم اطلاعاتی ذخیره می‌شوند و اجرای واقعی فرآیند را ضبط می‌کنند. تجزیه و تحلیل داده‌های رویداد، درک مفصلی از فرآیند را ایجاد می‌کند. اکثر تکنیک‌های فرآیندکاوی فرض می‌کنند که داده‌های رویداد از سطح ریزدانگی یکسان یا مناسب برخوردار هستند حال آن که چنین نیست.

نگاره‌رویدادهایی وجود دارند که جزئیات زیادی را در خود ذخیره می‌کنند بنابراین مدل‌هایی که از روی آن‌ها ساخته می‌شوند بسیار پیچیده هستند، از این رو به دنبال راه‌حلی هستیم که از جزئیات بکاهیم و آنها را به سطح بالاتری از انتزاع ببریم.



(الف)

(ب)

شکل (۱): الف) مدل استخراج شده از روی داده های خام مربوط به جریان کلیک ها (سطح پایین ریزدانگی). ب) مدل استراج شده از داده هایی که در سطوح بالاتری از انتزاع قرار گرفته اند (سطح بالای ریزدانگی) [2].

قسمت الف از شکل (۱) مدلی را نشان می‌دهد که به صورت خودکار از داده‌های خام «جریان کلیک» ها استخراج شده است. واضح است که تحلیل چنین مدلی بسیار سخت و طاقت‌فرسا بوده و عملاً نمی‌توان کار خاصی بر روی آن انجام داد. به عنوان مثال بررسی انطباق و بهبود فرآیند مشکل خواهد بود. در قسمت ب، انتزاع بر روی نگاره رویداد قبلی صورت گرفته و مدلی که استخراج شده قابل درک و تحلیل است.

به عبارت دقیق‌تر، در انتزاع رویدادها فرض می‌کنیم σ نگاره رویداد با جزئیات زیاد باشد. به دنبال تکنیکی مثل α هستیم که $\sigma' = \alpha(\sigma)$ که در آن σ' نگاره رویدادی با جزئیات کم است و به طور معمول باید داشته باشیم: $|\sigma| > |\alpha(\sigma)|$. به عبارت دیگر، باید پیچیدگی کاهش یافته باشد [2].

امروزه به دلیل پیشرفت تکنولوژی در سیستم‌های موقعیت‌یاب، موقعیت مکانی بسیاری از رویدادهایی که در طی یک فرآیند اتفاق می‌افتد ثبت می‌شوند. در فرآیندهایی مثل فرآیندهای دریافت و تحویل بسته، رویدادها در موقعیت‌های جغرافیایی مختلفی اتفاق می‌افتند. لازم است تا این نوع فرآیندها بر اساس تغییرات موقعیت مکانی فعالیت‌ها مورد تحلیل قرار گیرند. حال اگر تعداد موقعیت‌های مکانی زیاد باشد، در مدل‌سازی فرآیند با چالش مدل پیچیده مواجه خواهیم بود. در مقالات پیشین مرتبط با انتزاع رویدادها، کم‌تر به داده‌های جغرافیایی نگاره‌رویداد توجه شده است. از این رو راه‌حلی که در این مقاله ارائه می‌شود، به دنبال این است تا با بهره‌گیری از یک روش فضایی، بر اساس موقعیت مکانی، رویدادها به سطوح بالاتری از انتزاع برده شده و مشکل پیچیدگی مدل‌ها برطرف شود.

ساختار مقاله به شرح زیر است.

۲- مروری بر کارهای گذشته

در این بخش به تکنیک‌هایی که در کارهای پیشین برای انتزاع رویداد معرفی شده است خواهیم پرداخت. این تکنیک‌ها در سه دسته تکنیک‌های بدون ناظر، تکنیک‌های با ناظر و سایر تکنیک‌ها دسته‌بندی شده‌اند.

۲-۱- تکنیک‌های بدون ناظر

دسته‌ای از مقالات با استفاده از روش‌های بدون ناظر عمل انتزاع را انجام داده‌اند. اساس کار این تکنیک‌ها، تکرار زیاد در الگوهاست. تکرار در بعضی الگوها ساده و در برخی پیچیده است. در بعضی مقالات از تکنیک‌های خوشه-بندی برای انتزاع رویدادها استفاده شده است. در برخی دیگر، زیردنباله‌های پرتکرار و مرتبط پشت سر هم به یک فعالیت در سطح بالا نگاشت می‌شوند.

تکنیک دیگر، استفاده از قطعه قطعه کردن دنباله[1] است که در مرجع شماره [3] معرفی شده است. ایده‌ی کلی قطعه قطعه کردن دنباله این است که زیردنباله‌هایی که به نحوی مرتبط هستند، گروه‌بندی شوند. به طور خلاصه دستاورد این مقاله را در سه مرحله می‌توان بیان نمود: اول رویدادهای موجود بر اساس هم‌بستگی[2] بین‌شان در کلاس‌های مختلف قرار می‌گیرند. در مرحله‌ی بعد یک خوشه‌بندی سلسله‌مراتبی بر روی کلاس‌های رویدادها انجام می‌پذیرد و در آخر سطح دلخواهی از خوشه‌بندی برای انتزاع رویدادها انتخاب می‌شود.

در بعضی دیگر از مقالات از تکنیک‌های تعبیه کلمه استفاده شده است بدین صورت که کلمات با معنای مشابه در یک گروه قرار می‌گیرند [4]. استخراج مدل‌های پنهان مارکوف از طریف محاسبه احتمال نگاره‌ها و الگوریتم ویتربی نوع دیگر از تکنیک‌هاست که در حالت بدون ناظر استفاده شده است [5]. در برخی مقالات نیز مجاورت رویدادها برای انتزاع استفاده شده است که در آن بر اساس همزمانی اتفاق افتادن رویدادها، آن‌ها گروه‌بندی می‌شوند [6].

در مرجع شماره [6]، با استفاده از روش سلسله‌مراتبی خوشه‌بندی، سعی شده است تا اجزای مدل مرجع از نگاره رویداد استخراج شوند. مدل مرجع یک چارچوب انتزاعی یا یک هستی‌شناسی[3] حوزه‌ی خاص است که به طور شفاف روابط بین مفاهیم را بیان می‌کند. به عبارت دیگر، مدل‌های مرجع، مدل‌های

[1] Trace segmentation
[2] Correlation
[3] Ontology

مفهومی خاصی هستند که برای طراحی سایر مدل‌ها مورد استفاده مجدد قرار می‌گیرند. اجزای مدل مرجع هم بلاک‌هایی هستند که به طور فراوان در مدل‌ها ظاهر می‌شوند. ایده‌ی تکنیکی که در مرجع شماره [6] استفاده شده به این صورت است که فعالیت‌های موجود در رویدادها بر اساس اتفاق افتادن در مجاورت یکدیگر خوشه‌بندی شده و اجزای مدل مرجع از خوشه‌ها به دست می‌آیند. به عنوان مثال، برای دو فعالیت a و b، در همه دنباله‌هایی که شامل هر دو فعالیت هستند بررسی می‌شود که چند فعالیت در بین این دو وجود دارد؟ پس از شمارش تعداد آن‌ها، تعداد بر اندازه دنباله تقسیم می‌شود. این کار برای همه‌ی دنباله‌ها انجام می‌پذیرد و پس از محاسبه‌ی مجموع اعداد، حاصلجمع به تعداد دنباله‌هایی که شامل هر دو فعالیت a و b هستند تقسیم می‌شود تا به یک معیار نرمال شده برسیم. عددی که در نهایت به دست می‌آید، میزان مجاورت دو فعالیت را معلوم می‌کند. سپس با استفاده از تکنیک‌های سلسله مراتبی خوشه‌بندی فعالیت‌های نزدیک به هم در خوشه‌های یکسان قرار می‌گیرند.

تکنیک پیشنهادی توسط آقای مَنهارت و همکارش در [7] بر این اساس است که ابتدا مدل‌های فرآیند به صورت محلی استخراج شوند و با استفاده از این مدل‌ها، عمل انتزاع بر روی کل نگاره رویداد انجام شود. تکنیک ارائه شده چهار قسمت اصلی دارد: اول تعداد مشخصی از مدل های فرآیند محلی[1] توسط روش‌های اعلام شده در مرجع [7] استخراج می‌شود. در مرحله‌ی دوم به چالش هم‌پوشانی مدل‌های محلی پرداخته می‌شود. در این‌جا چالشی که وجود دارد این است که ممکن است مدل‌های محلی در برخی حالات هم‌پوشانی داشته باشند. برای حل این چالش به هریک از مدل‌های محلی یک امتیاز نسبت داده می‌شود. در مرحله‌ی سوم، هر یک از مدل‌های محلی به یک فعالیت سطح بالا نگاشت می‌شوند و در مرحله‌ی چهارم مدل‌سازی نهایی با استفاده از نگاره رویداد انتزاعی انجام می‌پذیرد.

در [8] که توسط آقای ژی‌لو و همکارانش ارائه شده است، نویسندگان از روش خود با نام FlexHMiner یاد کرده‌اند. پیاده‌سازی این روش نیز در افزونه‌ای با همین نام برای ابزار ProM انجام شده است. FlexHMiner از سه گام اصلی تشکیل شده است: محاسبه‌ی درخت فعالیت‌ها، محاسبه‌ی نگاره‌های انتزاعی و محاسبه‌ی مدل‌ها. برای محاسبه‌ی درخت فعالیت‌ها، سه روش متفاوت بیان شده است. در حالت اول از دانش حوزه، استفاده از خوشه‌بندی تصادفی و استفاده از درخت هموار. برای محاسبه‌ی نگاره‌های انتزاعی و مدل‌سازی، رفتار زیرفرآیندها مورد تحلیل قرار گرفته و به سطوح بالاتر نگاشت می‌شود و در نهایت مدل‌ها از سطح بالا استخراج می‌شوند. برای ارزیابی این مقاله از چهار معیار که بر روی هفت مجموعه داده اندازه‌گیری شده، استفاده شده است، مناسب بودن، دقت، معیار F و پیچیدگی. برای محاسبه‌ی اندازه پیچیدگی، با فرض مدل‌سازی با شبکه پتری، مجموع فعالیت‌ها و کمان‌ها قبل و بعد از انتزاع محاسبه می‌شود.

۲-۲- تکنیک‌های با ناظر

تکنیک‌های با ناظر معمولاً سه حالت دارند. در دسته‌ی اول، از بازه‌های زمانی برای گروه‌بندی رویدادها استفاده می‌شود. در دسته‌ی دوم، قسمت کوچکی از نگاره رویدادها انتخاب و به صورت دستی نگاشت می‌شوند. سپس همین نگاشت کوچک به عنوان پایگاه دانش داده‌های بزرگ‌تر لحاظ می‌شود. در دسته‌ی سوم نیز از مدل‌های مرجع استفاده می‌شود [2].

تکنیک ارائه شده توسط آقای دی‌لئونی و همکارش در [9] بر این اساس است که ابتدا رویدادهای موجود در یک دنباله در نشست[2]هایی قرار می‌گیرند و با خوشه‌بندی این نشست‌ها، فعالیت‌های سطح بالا مشخص می‌شوند. نحوه‌ی ایجاد نشست به این صورت است که دنباله $\sigma = \langle e_1 . e_2 . \dots . e_n \rangle$ به نشست‌های $s_\Delta(\sigma) = \langle s_1 . \dots . s_m \rangle$ تقسیم‌بندی می‌شود با این شرایط که اولاً فاصله‌ی زمانی بین شروع یک نشست و پایان نشست قبلی از یک حد آستانه (Δ) بیشتر باشد، دوماً فاصله زمانی بین هر دو رویداد واقع در یک نشست کمتر از حد آستانه (Δ) باشد. پس از این مرحله، نشست‌ها برای عمل خوشه‌بندی در بردارهایی کدگذاری[3] می‌شوند.

نگاشت مسیرهای رفت و آمد مشتریان (CJM)[4] یک تکنیک پرطرفدار در سازمان‌هاست تا بتوانند رفتار مشتریان خود را برای تحلیل اوضاع و فروش بیشتر درک نمایند. کم‌ترین منفعت استفاده از CJMها درک مسیرهای اصلی تردد مشتریان است [10]. حال اگر با مدل مواجه شویم که مسیرهای تردد مشتریان را با جزئیات زیاد نشان داده باشد، سودی در تحلیل به دست نخواهیم آورد. در مرجع [10] یک ابزار مبتنی بر جاوااسکریپت معرفی شده است که مدل‌ها را در سطوح مختلف ریزدانگی نشان می‌دهد. تکنیک ارائه شده در این مقاله شامل چهار مرحله است. در مرحله اول درخت‌های فرآیند از نگاره رویداد استخراج می‌شوند. درخت فرآیند یک بازنمایی انتزاعی سلسله‌مراتبی از مدل فرآیند است. مرحله دوم مربوط به کشف مسیرهای تردد مشتریان است. در این مرحله CJMها از نگاره رویداد استخراج می‌شوند. در مرحله سوم با تعاملی که با کاربر نهایی صورت می‌گیرد، فعالیت‌ها در یکدیگر ادغام می‌شوند و در مرحله چهارم مسیرهای رفت و آمد در سطوح مختلف ریزدانگی قرار می‌گیرند.

در مرجع شماره [11]، یک روش مبتنی بر الگوهای رفتاری فعالیت‌ها ارائه شده است. فرض این روش بر این است که دانش حوزه در دسترس است. در ابتدا ابعاد مختلف فعالیت‌ها شناسایی می‌شود، این کار به صورت دستی و یا ماشینی انجام می‌گیرد و سپس مدل‌های انتزاعی استخراج می‌شوند. پس از کشف مدل، یک مرحله ترازبندی بین مدل‌ها و رویدادهای سطح پایین انجام می‌پذیرد تا صحت کار تایید شود.

در بعضی از تکنیک‌های با ناظر، کاربر باید یک مدل ماکرو (که یک بازنمایی از مدل مارکوف است) ورودی را به تکنیک بدهد و همچنین فعالیت‌های سطح بالا و روابط بین آن‌ها را بشناسد. در بعضی تکنیک‌های دیگر، باید توضیحات کافی در نگاره رویدادها برای انتزاع وجود داشته باشد. برای دسته‌بندی داده‌های پیوسته، نظیر داده‌های سنسورها، یک اندازه پنجره با اندازه مینیمال در نظر گرفته می‌شود. سپس این دسته‌ها خوشه‌بندی شده و بر اساس ویژگی‌های موجود در آن‌ها یک برچسب مناسب به هر خوشه تعلق می‌گیرد. از دیگر موارد تکنیک‌های با ناظر، می‌توان به مواردی همچون استفاده از آنتولوژی، حل مسائل ارضای محدودیت، استفاده از توزیع احتمال، مدل پنهان مارکوف و کشف الگوهای رفتاری بر اساس شبکه پتری اطلاعات اشاره نمود [2].

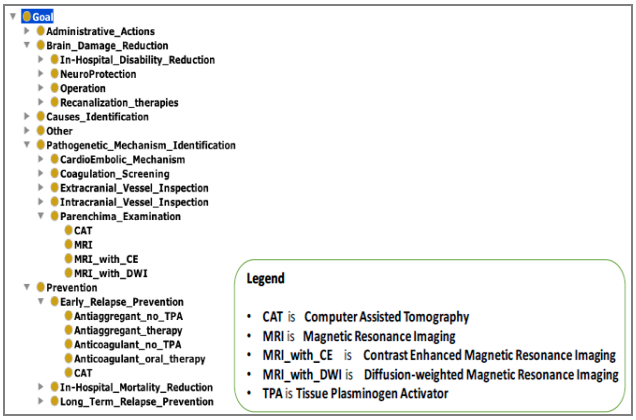
[1] Local Process Model (LPM)

[2] Session
[3] Encode
[4] Customer Journey Mapping

۲-۳- سایر تکنیک‌ها

در برخی مقالات از روش‌های معنایی برای انتزاع رویداد استفاده شده است. می‌دانیم که با استفاده از آنتولوژی می‌توان با مشخص کردن مفاهیم و موجودیت‌های یک حوزه، توصیف روابط بین آن‌ها و به خدمت گرفتن مجموعه‌ای از قواعد، حوزه را به صورت رسمی و قابل فهم برای ماشین بیان کرد. به عبارت دیگر، در یک دامنه یا حوزه‌ی خاص، تمام اشیا یا موجودیت‌هایی که وجود دارند و یا ممکن است وجود داشته باشند مورد مطالعه قرار گرفته و بر اساس ویژگی‌هایشان دسته‌بندی می‌شوند و ارتباط بین اشیا نیز مشخص می‌شود.

در مرجع شماره [12] از روش‌های معنایی برای انتزاع رویدادها استفاده شده است. همان‌طور که در شکل (۲) نشان داده شده، با دانشی که از حوزه‌ی پزشکی وجود دارد، رویدادهای سطح پایین در سطوح بالاتر تجمیع می‌شوند. مثلاً MRI و CAT در فعالیت سطح بالاتر که Parenchima _Examination است ادغام می‌شوند.



شکل (۲) : انتزاع بر اساس هستی‌شناسی حوزه [12]

بعضی از مقالاتی که به انتزاع داده‌های موقعیت مکانی پرداخته‌اند، مربوط به داده‌های سنسور هستند که غالباً موقعیت مکانی درون ساختمان را در نظر گرفته‌اند. در مرجع شماره [13]، از تکنیکی به نام ROAD استفاده شده است. مجموعه داده‌ای که در این مقاله از آن بهره گرفته شده، مربوط به بیمارستان بیماران سرطانی‌ست که به صورت سرپایی خدماتی نظیر شیمی درمانی را دریافت می‌کنند. به هر یک از موجودیت‌ها (بیماران، پزشکان و پرستاران) یک سنسور متصل است که هر ۳ ثانیه موقعیت مکانی آن‌ها را ثبت می‌کند. در این تکنیک، به دلیل این‌که رویدادها در فواصل زمانی مشخص ثبت می‌شوند، در مرحله‌ی پیش‌پردازش یک مرحله تجمیع انجام می‌گیرد به این صورت که در رویدادهایی که پس از گذشت زمان موقعیت مکانی آن‌ها تغییر نمی‌کند، این رویدادها با یکدیگر تجمیع می‌شوند. این تکنیک شامل دو مرحله است. مرحله اول تکنیک ROAD، کاوش تعاملات بین موجودیت‌ها در موقعیت‌های مکانی مختلف است. مثلاً اگر در یک فرآیند بیمارستانی، یک بیمار، یک پرستار و یک پزشک در حدود ساعت ۱۴ در یک موقعیت مکانی یکسان قرار گیرند، فعالیت که به سطح بالا نگاشت می‌شود، «معاینه» خواهد بود. در مرحله دوم این تکنیک، با استفاده از روش برنامه‌ریزی صحیح و به کارگیری دانشی که از فرآیند وجود دارد، تطابق بهینه صورت می‌گیرد.

۳- مفاهیم پایه

در این بخش به طور خلاصه مفاهیم پایه‌ای حوزه معرفی می‌شوند:

- فرآیندکاوی: موضوعی است که بین هوش محاسباتی و داده‌کاوی از یک سو و مدل‌سازی و تحلیل فرآیندها از سوی دیگر قرار دارد و هدف از آن کشف، نظارت و بهبود فرآیندهای واقعی است.
- نگاره رویداد: رویدادهایی که در فرآیندها رخ می‌دهند، به همراه ویژگی‌هایشان نظیر شناسه رویداد ، شناسه نمونه ، زمان شروع و پایان رویداد، فعالیت‌ها و ... در لیستی ثبت می‌شوند که نقطه‌ی شروع فرآیندکاوی‌ست و نگاره رویداد نام دارد.
- فعالیت: عبارت است از عملی که در زمان مشخصی از فرآیند توسط عامل انسانی یا ماشینی انجام می‌پذیرد.
- دنباله رویداد: مجموعه‌ای از رویدادها که در اجرای یک فرآیند اتفاق می‌افتند.
- ریزدانگی رویدادها: منظور از ریزدانگی رویدادها، سطح جزئیاتی است که از رویدادها ثبت شده است. هر چه جزئیات بیشتری از رویدادها ثبت شده باشد، میزان ریزدانگی بیشتر است.
- انتزاع رویدادها: تکنیک‌هایی که بر روی رویدادهای با سطح ریزدانگی زیاد اعمال می‌شود تا از جزئیات کاسته شود و در نتیجه‌ی آن مدل‌های قابل فهم‌تری استخراج شود.
- طول و عرض جغرافیایی: با استفاده از طول و عرض جغرافیایی می‌توان هر نقطه بر روی کره‌ی زمین را به صورت منحصربه‌فرد مشخص نمود. طول جغرافیایی عددی بین صفر و ۱۸۰ و عرض جغرافیایی عددی بین صفر و ۹۰ است.

۴- بیان مسئله

در فرآیندهایی مثل فرآیندهای دریافت و تحویل بسته، رویدادها در موقعیت‌های جغرافیایی مختلفی اتفاق می‌افتند. لازم است تا این نوع فرآیندها بر اساس تغییرات موقعیت مکانی فعالیت‌ها مورد تحلیل قرار گیرند. حال اگر تعداد موقعیت‌های مکانی زیاد باشد، در مدل‌سازی فرآیند با چالش مدل پیچیده مواجه خواهیم بود. در مقالات پیشین برای انتزاع رویدادها کم‌تر به داده‌های جغرافیایی نگاره‌رویداد توجه شده است. از این رو راه‌حلی که در این‌جا ارائه می‌شود، به دنبال این است تا با بهره‌گیری از یک روش فضایی، بر اساس موقعیت مکانی، رویدادها به سطوح بالاتری از انتزاع برده شده و مشکل پیچیدگی مدل‌ها برطرف شود.

هدف از این پژوهش، ارائه‌ی یک رویکرد فضایی برای انتزاع رویدادهاست. به عبارت دقیق‌تر، قصد داریم تا با به‌کارگیری خوشه‌بندی سلسله‌مراتبی بر روی موقعیت جغرافیایی رویدادها، به نوعی آن‌ها را در سطوح بالاتر گروه‌بندی کنیم تا تحلیل بهتر آن‌ها میسر شود. با در اختیار داشتن فراداده‌های مناسب نظیر اطلاعات شهرها و استان‌ها می‌توان سطوح انتزاع را معنادارتر تعریف نمود و مدل را مثلاً از لحاظ تغییرات موقعیت جغرافیایی در سطح شهر، استان، ایالت، کشور و ... ترسیم نمود.

۵- سیستم پیشنهادی

۵-۱- جریان کاری سیستم پیشنهادی

جریان کاری سیستم پیشنهادی در شکل (۳) نشان داده شده است، رویکرد این پژوهش شامل ۳ فاز است. هر بخش دارای یک بخش ورودی، یک بخش عملیات پردازشی و یک بخش خروجی است، بدین صورت که خروجی هر بخش به عنوان ورودی بخش دیگر در نظر گرفته می‌شود. فاز یک مربوط به پیش‌پردازش نگاره‌رویداد است. نگاره‌رویدادها توسط سیستم‌های اطلاعاتی از فرآیندهای واقعی استخراج می‌شوند. در فاز دوم، عمل انتزاع بر روی رویدادها انجام شده و رویدادهای انتزاعی در سطوح مختلف تجمیع می‌گردند. فاز سوم نیز مربوط به مدل‌سازی و ارزیابی کار است که البته مصورسازی مدل‌ها در حیطه‌ی این پژوهش نیست و در ادامه‌ی کار ما تعریف می‌شود.

۵-۱-۱- فاز اول: پیش‌پردازش نگاره‌رویداد

در مرحله اول، پیش‌پردازش‌های لازم بر روی نگاره رویداد انجام می‌شود. به عنوان نمونه فیلدهای غیر لازم از نگاره رویداد کنار گذاشته می‌شوند. همچنین نمونه‌هایی که کامل نیستند از نگاره رویداد حذف می‌شوند، مثلاً نمونه‌هایی که قبل از شروع ثبت رویدادها آغاز شده و نیمی از آن‌ها در نگاره رویداد وجود دارد و یا نمونه‌هایی که در انتهای نگاره رویدادکامل نشده‌اند. از آن‌جایی که کار اصلی پژوهش نیاز به داده‌های جغرافیایی دارد، اگر این داده‌ها در مجموعه داده حضور نداشتند، لازم است تا نگاره رویداد با داده‌های جغرافیایی غنی‌سازی شود.

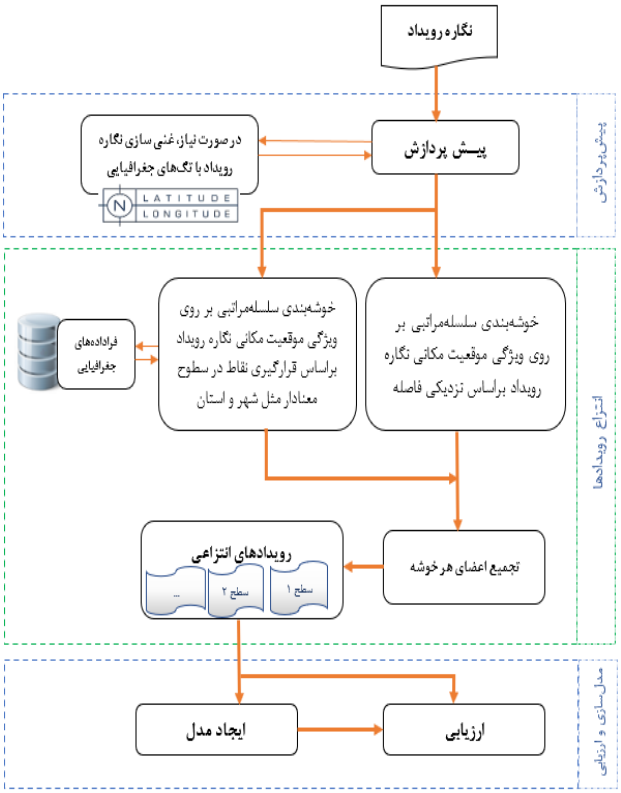
این داده‌ها شامل طول و عرض جغرافیایی است. می‌دانیم طول جغرافیایی عددی بین صفر و ۱۸۰ و عرض جغرافیایی عددی بین صفر و ۹۰ است که واقع شدن در نیم‌کره‌ی شمالی یا جنوبی و همچنین نیم‌کره‌ی شرقی یا غربی این اعداد را مثبت یا منفی می‌کند. هر نقطه بر روی کره‌ی زمین به طور منحصر به فرد با این دو عدد قابل مشخص کردن است.

۵-۱-۲- فاز دوم: انتزاع رویدادها

در مرحله دوم، ابتدا نگاره‌رویداد با توجه به ویژگی موقعیت مکانی که در مرحله‌ی قبل به آن اضافه شده است، بر اساس معیار فاصله و با دو رویکرد خوشه‌بندی می‌شود. ایده کلی استفاده از روش خوشه‌بندی سلسله مراتبی‌ست تا انتزاع رویدادها در سطوح مختلف فراهم می‌شود. در رویکرد اول، فراداده‌های مرتبط در دسترس نیست و نگاره‌رویداد صرفاً بر اساس نزدیکی فاصله نقاط از یکدیگر خوشه‌بندی می‌شوند. برای معیار فاصله، می‌توان فاصله اقلیدسی بین دو نقطه و یا فاصله‌ی راه‌های عبور و مرور را در نظر گرفت. در رویکرد دوم، با استفاده از فراداده‌های جغرافیایی، خوشه‌بندی دقیق‌تر می‌شود بدین صورت که به ازای هر نقطه تعیین می‌شود که آن نقطه متعلق به کدام شهر و کدام استان است. ابزار Open Street Map به عنوان ابزار مورد نیاز برای تأمین فراداده‌های کمکی انتخاب شده است. پس از انجام خوشه‌بندی، داده‌هایی که در یک خوشه قرار می‌گیرند با یکدیگر تجمیع شده و یک نماینده از آن‌ها در داده‌های انتزاعی حضور خواهد داشت.

۵-۱-۳- فاز سوم: مدل‌سازی و ارزیابی

مرحله سوم نیز مربوط به مدل‌سازی فرآیند با داده‌های انتزاعی و ارزیابی کار است. بهتر است مدل‌ها بر روی نقشه‌ی واقعی نمایش داده شوند تا تاثیر انتزاع رویدادها در مدل بهتر نمایان شود، البته مصورسازی جزء اصلی کار ما به شمار نمی‌آید و در ادامه‌ی کار ماست. در انتها نیز نتایج حاصل از پژوهش مورد ارزیابی قرار خواهد گرفت. برای ارزیابی پژوهش، مجموعه داده شرکت ملی پست ایران که در دسترس است در نظر گرفته شده است.



شکل (۳): جریان کاری سیستم پیشنهادی

۵-۲- پیاده‌سازی

۵-۲-۱- مجموعه داده

مجموعه داده ای که در این تحقیق مورد استفاده قرار گرفته است، مربوط به فرایند دریافت و تحویل بسته در یک شرکت پستی است که در بازه زمانی ماه ژوئن سال ۲۰۱۷ جمع آوری شده است.

در جدول (۱) ویژگی‌های فایل رویداد فرآیند دریافت و تحویل بسته نشان داده شده است. مجموعه داده‌ی انتخاب شده، شامل ۲۶۱۷۳۲۴ رکورد منحصربه‌فرد است.

جدول (۱): ویژگی‌های فایل رویداد مجموعه داده شرکت پست

عنوان ستون	توضیح
parcel_code	شناسه بسته پستی
post_node_id	شناسه ایستگاه پستی
event_date	زمان وقوع رویداد
lat	عرض جغرافیایی
lng	طول جغرافیایی

۵-۲-۲- پیاده‌سازی رویکرد اول: استفاده از فراداده‌ها

به منظور در اختیار داشتن فراداده‌های کمکی، در این پژوهش از Open Street Map[1] استفاده شده است. با استفاده از OSM این امکان وجود دارد تا هر نقطه‌ی جغرافیایی که در نگاره‌رویداد وجود دارد، به یک شهر یا استان نگاشت شود[2].

قابلیتی در OSM به نام Overpass API وجود دارد که می‌توان با اعمال پرس‌وجوهای مناسب داده‌های دلخواه را از آن استخراج نمود. به عنوان مثال، پرس‌وجوی شماره (۱) داده‌های محدوده‌ای مربوط به استان‌های ایران را هم به صورت نقشه و هم به صورت داده در اختیار قرار می‌دهد.

```
[out:xml][timeout:500][bbox:28.5,47,38,60];
//gather results
(
    rel[admin_level=6]                    ( پرس‌وجوی شماره ۱ )
       [type=boundary]
       [boundary=administrative];
);
//print results
out meta;
>;
out skel qt;
```

در این پژوهش با استفاده از Overpass API و همچنین ابزار برخط دیگری به نشانی http://polygons.openstreetmap.fr/index.py، داده‌های محدوده‌ای مربوط به هر شهر و استان کشور ایران جمع‌آوری شد. سپس با استفاده از ابزار JTS Topology Suite واسطی طراحی گردید که هر نقطه‌ی دلخواه بر روی نقشه ایران را به یک شهر و یک استان نگاشت می‌کند.

۵-۲-۳- پیاده‌سازی رویکرد دوم: خوشه‌بندی سلسله‌مراتبی نقاط جغرافیایی

مطابق آن‌چه در مرجع شماره [14] آمده است، روش‌های پیوند گوناگونی برای الگوریتم‌های خوشه‌بندی ذکر شده است مثل پیوند تکی، پیوند میانگین و پیوند کامل. معیارهای مختلفی را می‌توان برای اندازه‌گیری فاصله بین خوشه‌ها به کار برد. برای مثال می‌توان فاصله را براساس فاصله بین نزدی‌کترین یا دورترین مشاهدات بین دو خوشه در نظر گرفت. هر یک از این معیارها معایب و مزایای مخصوص به خود را دارند. ولی با توجه به ساختار داده‌ها (وجود داده‌های پرت، الگوی پراکندگی مشاهدات در هر خوشه و...) می‌توان مبنای انتخاب یکی از روش‌های پیوند باشد.

در آزمایشات انجام شده با روش‌های پیوند مختلف و هم‌چنین تعداد خوشه‌های متفاوت، دو نتیجه‌ی زیر به دست آمد (اثبات این دو نتیجه در بخش ارزیابی ذکر می‌شود):

- بهترین تعداد خوشه برای خوشه‌بندی سلسله‌مراتبی، ۱۳ خوشه است.
- در مجموعه داده شرکت پست، استفاده از روش‌های مختلف پیوند خروجی یکسان خواهد داشت.

در این روش، مراکز پستی بر اساس فاصله جغرافیایی خوشه‌بندی شده و در هر خوشه، مرکزی به عنوان نماینده آن خوشه قرار می‌گیرد که از نظر زمان دریافت و ارسال بسته نسبت به سایر مراکز کمینه باشد. سناریویی را فرض کنید که یک بسته پستی از شهر گناباد از استان خراسان‌رضوی به شهر بیرجند از استان خراسان‌جنوبی ارسال می‌شود. اگر طبق روش انتزاع اول عمل شود که همین روش در ارسال مرسولات پستی نیز انجام می‌گیرد، این بسته ابتدا از گناباد به مشهد که مرکز استان خراسان‌رضوی است هدایت شده و سپس از مشهد به بیرجند ارسال خواهد شد. در صورتی‌که فاصله شهر گناباد تا بیرجند از فاصله گناباد تا مشهد به مراتب کمتر است. در این وضعیت، روش دوم که بر اساس فاصله، خوشه‌بندی را انجام می‌دهد می‌تواند بسته پستی را از گناباد مستقیماً به بیرجند ارسال نماید، بدون آن‌که این بسته به مشهد هدایت شود. لذا در سناریوهایی که یکی از شهرهای مبدا و مقصد مرسوله پستی در مرز یک استان قرار گرفته باشد، روش رایج برای ارسال بسته ناکارا بوده و زمان و هزینه بیشتری را تحمیل خواهد کرد، در حالی‌که در روش خوشه‌بندی سلسله‌مراتبی این ایراد وجود ندارد.

پیاده‌سازی طرح پیشنهادی در این نوشتار با زبان جاوا انجام شده است. به منظور پیاده‌سازی خوشه‌بندی سلسله‌مراتبی، از کتابخانه Weka که یک کتابخانه شناخته شده در حوزه‌ی کارهای یادگیری ماشین است، استفاده کردیم.

۵-۳- ارزیابی

برای ارزیابی انتخاب بهترین تعداد خوشه، به ازای تعداد خوشه ۷ الی ۱۶، با چهار روش مختلف پیوند، آزمایشات تکرار شد و در جدول (۲) نتایج آورده شده است. با توجه به این نتایج، تعداد ۱۳ خوشه به عنوان بهترین حالت خوشه‌بندی انتخاب شد. رابطه (۱) و (۲) نحوه محاسبه‌ی میانگین مجموع فاصله‌ی نقاط فرآیند خوشه‌بندی به ازای تعداد خوشه‌های مختلف را نشان می‌دهند.

- (lon, lat): مشخصات نقاط موجود در یک خوشه
- مجموع فاصله‌ی نقاط هر خوشه : A
- میانگین مجموع فاصله‌ی نقاط کل فرآیند خوشه‌بندی : B
- تعداد خوشه‌ها: n

$$A = \sum \sqrt{(lon_i - lon_j)^2 + (lat_i - lat_j)^2} \quad (۱)$$

$$B = \frac{1}{n} . \sum A \quad (۲)$$

[1] https://www.openstreetmap.org/
[2] https://wiki.openstreetmap.org/wiki/Main_Page

جدول (۲): یافتن بهترین تعداد خوشه بر اساس فاصله‌ی نقاط هر خوشه از یکدیگر

	میانگین مجموع فاصله نقاط در خوشه (B)			
	Single Linkage	Complete Linkage	Average Linkage	Centroid
n=7	1.87588	1.87588	1.87588	1.87588
n=8	1.64139	1.64139	1.64139	1.64139
n=9	1.45902	1.45902	1.45902	1.45902
n=10	1.31311	1.31311	1.31311	1.31311
n=11	1.193274	1.193274	1.193274	1.193274
n=12	1.09426	1.09426	1.09426	1.09426
n=13	**1.01009**	**1.01009**	**1.01009**	**1.01009**
n=14	9.37942	9.37942	9.37942	9.37942
n=15	8.75413	8.75413	8.75413	8.75413
n=16	8.20699	8.20699	8.20699	8.20699

معیار دیگری که در مراجعی مثل مرجع [8] از آن استفاده شده است، معیار پیچیدگی است. این معیار اگر چه شاید به تنهایی برای ارزیابی مناسب نباشد، اما به صورت کمی میزان کاهش پیچیدگی را بیان می‌کند. در ادامه روابط مربوطه نشان داده شده است.

- M: مدل استخراج شده قبل از انتزاع رویدادها
- M’: مدل استخراج شده بعد از انتزاع رویدادها
- a: تعداد گره‌های مدل M
- b: تعداد گره‌های مدل M’

$$\begin{cases} complexity_{size}(M) = 1 \\ complexity_{size}(M') = {}^{b}/_{a} \end{cases} \quad (۳)$$

در جدول شماره (۳) نتایج مربوط به معیار پیچیدگی به ازای تعداد خوشه‌های مختلف نشان داده شده است.

جدول (۳): مقایسه پیچیدگی قبل و بعد از انتزاع رویدادها

	پیچیدگی قبل از انتزاع	پیچیدگی بعد از انتزاع
n=7	1	0.01781
n=8	1	0.02035
n=9	1	0.02290
n=10	1	0.02544
n=11	1	0.02796
n=12	1	0.03053
n=13	1	0.03307
n=14	1	0.03562
n=15	1	0.03816
n=16	1	0.04071

در انتهای بخش ارزیابی، نگاهی تحلیلی داریم به مقایسه‌ی نتایج دو رویکرد انتزاع. در ابتدا مفهومی به نام هم‌پوشانی را تشریح می‌کنیم. یک نقطه جغرافیایی در رویکرد خوشه‌بندی با رویکرد فراداده هم‌پوشانی دارد اگر با مرکز استان خود در یک خوشه قرار گیرد.

از مفهوم هم‌پوشانی برای شناسایی حالاتی استفاده می‌کنیم که خوشه‌بندی نقاط عملکرد بهتری نسبت به استفاده از فراداده دارد، به عنوان مثال در ارسال بسته‌های پستی، ممکن است در برخی نقاط مرزی استان‌ها بهتر باشد بسته پستی به جای ارسال به مرکز استان، به مرکز استان مجاور ارسال شود. در ادامه تحلیل‌های مربوط به این بخش از نتایج آزمایشات به‌دست آمده است، ذکر شده است.

- به طور متوسط ۱۵٪ بین دو رویکرد هم‌پوشانی وجود ندارد.
- طبق آزمایشات، ۲۹ نقطه پستی به ازای تعداد خوشه‌های مختلف، با مرکز استان خود هم‌خوشه نشده‌اند. مثال: نقطه پستی شهربابک در استان کرمان.
- غیرهم‌پوشانی‌ها بیشتر در نقاط مرزی استان‌ها مشاهده می‌شود (همه‌ی ۲۹ مرکز پستی در مرز استان بوده‌اند).
- طبق آزمایش‌های به عمل آمده، به طور متوسط به ازای تعداد خوشه‌های مختلف، 39.15% از غیرهم‌پوشانی‌ها در ۶ استان پهناورتر روی داده است (اصفهان، فارس، خراسان رضوی، یزد، کرمان، سمنان).
- به طور متوسط 25.2% از غیرهم‌پوشانی‌ها در استان‌های با تراکم شهری بالا مثل شهرهای شمالی روی داده است.
- استفاده از خوشه‌بندی بر اساس فاصله، در استان‌های پهناور و پرتراکم میتواند کارآیی فرآیندها را بهبود بخشد.

۶- نتیجه‌گیری

خروجی نگاره‌رویدادهای با جزئیات زیاد مدل‌هایی پیچیده و غیرقابل درک هستند. بنابراین باید تکنیک‌هایی را به کار گرفت تا با استفاده از آن‌ها بتوان از جزئیات کاست. در کارهای پژوهشی گذشته کم‌تر به جنبه‌ی موقعیت مکانی وقوع رویدادها توجه شده است. اگر بخواهیم فرآیندها را بر اساس تغییرات موقعیت مکانی مدل‌سازی نماییم، طبیعتاً جزئیات زیاد ما را به سمت مدل‌های پیچیده رهنمون می‌سازد.

ما در این کار پژوهشی از داده‌های موقعیت مکانی رویدادها استفاده کرده و با دو رویکرد به انتزاع رویدادها پرداخته‌ایم. در رویکرد نخست، از فراداده‌های کمکی استفاده شد تا انتزاع در سطوح معنادار مثل سطح شهر و سطح استان انجام پذیرد که این روش رایج در ارسال مرسولات در شرکت ملی پست است.

در رویکرد دوم، از الگوریتم سلسله‌مراتبی خوشه‌بندی استفاده شد و نقاط بر اساس نزدیکی فاصله‌شان از یکدیگر در خوشه‌های مربوطه قرار گرفتند. دلیل انتخاب خوشه‌بندی سلسله‌مراتبی از بین روش‌های دیگر مثل K-Means، انتزاع رویدادها در سطوح مختلف است. پس از عمل خوشه‌بندی، داده‌های انتزاعی با یکدیگر تجمیع شده و در سطوح بالاتر، یک نماینده از آن‌ها که زمان ارسال و دریافت بسته در آن کمینه است، حاضر می‌شود. لذا این روش می‌تواند در سناریوهایی که یکی از نقاط مبدا یا مقصد در مرز استان قرار دارند، به مراتب عملکرد بهتری نسبت به روش اول داشته باشد، لذا می تواند به عنوان روش جدید در ارسال مرسولات در شرکت ملی پست استفاده شود. در نتیجه خوشه بندی سلسله‌مراتبی می‌تواند کارآیی فرآیندها را بهبود ببخشد.

مراجع

[1] C. dos Santos Garcia, A. Meincheim, . E. Ribeiro Faria Junior, M. Rosano Dallagassa, . D. Maria Vecino Sato, D.

Ribeiro Carvalho, E. Alves Portela Santos and E. Emilio Scalabrin, "Process mining techniques and applications – A systematic mapping study," *Expert Systems with Applications*, vol. 133, pp. 260-295, 2019.

[2] S. J. van Zelst, F. Mannhardt, M. d. Leoni and A. Koschmider, "Event abstraction in process mining: literature review and taxonomy," *granular computing*, vol. 5, pp. 1-18, 2020.

[3] C. W. Günther, A. Rozinat and W. M. Van Der Aalst, "Activity mining by global trace segmentation," in *International Conference on Business Process Management*, 2009.

[4] D. Sánchez-Charles, J. Carmona, V. Muntés-Mulero and M. Solé, "Reducing event variability in logs by clustering of word embeddings," in *International Conference on Business Process Management*, 2017.

[5] A. Alharbi, A. Bulpitt and O. A. Johnson, "Towards unsupervised detection of process models in healthcare," 2018.

[6] J.-R. Rehse and P. Fettke, "Clustering business process activities for identifying reference model components," in *International Conference on Business Process Management*, 2018.

[7] F. Mannhardt and N. Tax, "Unsupervised event abstraction using pattern abstraction and local process models," in *Joint Proceedings of (EMISA) co-located with the 29th International Conference on Advanced Information Systems Engineering (CAiSE)*, 2017.

[8] L. Xixi, A. Gal and H. Reijers, "Discovering Hierarchical Processes Using Flexible Activity Trees for Event Abstraction," in *2nd International Conference on Process Mining (ICPM)*, 2020.

[9] M. de Leoni and S. Dündar, "Event-log abstraction using batch session identification and clustering," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020.

[10] B. Gaël and P. Andritsos, "Cjm-ab: Abstracting customer journey maps using process mining," in *International Conference on Advanced Information Systems Engineering*, 2018.

[11] F. Mannhardt, M. d. Leoni, H. A. Reijers, W. M. v. d. Aalst and P. J. Toussaint, "Guided process discovery - A pattern-based approach," in *Information Systems 76*, 2018.

[12] G. Leonardi, M. Striani, S. Quaglini, A. Cavallini and S. Montani, "Towards Semantic Process Mining Through Knowledge-Based Trace Abstraction," in *International Symposium on Data-Driven Process Discovery and Analysis*, 2017.

[13] A. Senderovich, . A. Rogge-Solti, A. Gal, J. Mendling and A. Mandelbaum, "The ROAD from sensor data to process instances via interaction mining," in *International Conference on Advanced Information Systems Engineering*, 2016.

[14] D. M. Christopher, R. Prabhakar and S. C. H. U. T. Z. E. Hinrich, Introduction to Information Retrieval, 2008.