

Minería de Procesos aplicada a la distribución postal

Victor Martinez¹, Laura Lanzarini^{1,2}, and Franco Ronchetti^{1,2,3}

¹ *Facultad de Informática, Universidad Nacional de La Plata*

² *Instituto de Investigación en Informática LIDI (UNLP-CIC)*

³ *Comisión de Investigaciones Científicas de la provincia de Buenos Aires (CIC-PBA)*

`martinezvictor@hotmail.com, {laural, fronchetti}@lidi.info.unlp.edu.ar`

Resumen La minería de procesos es una técnica que permite analizar procesos de negocio a través de registros de eventos. En este artículo se utilizan distintas técnicas de Minería de Procesos para analizar datos basados en la distribución postal de productos en la República Argentina durante los años 2017 a 2019. Los resultados obtenidos permiten afirmar que el 85 % de los envíos realizados se ajustan exactamente al modelo. El análisis de las situaciones con bajo nivel de ajuste al proceso descubierto constituyó una herramienta de identificación rápida de algunos problemas recurrentes en la distribución facilitando el análisis de los desvíos ocurridos. A futuro se espera poder incorporar estas técnicas en la construcción de alertas tempranas que adviertan de la existencia de desvíos excesivos del proceso.

Keywords: Process Mining, Data Mining, Postal Distribution, Postal Processes, Business Process Management

1. Introduction

En la actualidad la mayoría de los sistemas de información dejan registro de las actividades realizadas, ya sean sistemas de gestión empresarial (ERP, CRM, WMS, BI, etc) como también desarrollos propios de cada empresa. Por lo general se registra qué se realizó, quién lo realizó y cuándo se realizó, entre otras cosas. A partir de estos datos almacenados puede obtenerse información que describa el proceso y buscar mejoras o solucionar problemas.

Las técnicas de Minería de Procesos se utilizan para analizar estos datos y encontrar patrones de comportamiento. Se puede encontrar con que tareas comienza un proceso, cual es la secuencia que se cumple, y con que tareas termina. De esta forma se puede “descubrir” el proceso que se está realizando para una determinada actividad.

Con la minería de procesos se pueden responder preguntas como: ¿qué fue lo que pasó realmente?, ¿por qué pasó?, ¿qué podría pasar en el futuro?, ¿cómo se puede mejorar el control del proceso?, ¿cómo se puede rediseñar el proceso para mejorar la performance? [1].

Una de sus principales ventajas es que permite trabajar directamente sobre datos reales y obtener el verdadero comportamiento del proceso, el cual, en algunos casos, no es el diseñado originalmente.

En este trabajo se aplicarán las técnicas de minería de procesos a la distribución postal en la República Argentina para analizar su funcionamiento con el objetivo de encontrar desvíos operativos, cuellos de botella y otros problemas que impactan negativamente en la calidad del servicio.

En la distribución postal se lleva un registro de todas las actividades, desde el ingreso del producto hasta su entrega al cliente. Entre estas actividades se encuentra la recepción del envío, ingreso a un almacén, traslados internos o intentos de entrega, entre otras. Todas deben realizarse en un orden y en un intervalo de tiempo específicos. Ocasionalmente se producen desvíos, estos pueden ser redundancia o inconsistencia de tareas (se repiten o no se realizan en el orden correspondiente), excesivo tiempo para su finalización u otras.

En la literatura se pueden encontrar trabajos que relacionan el negocio postal, la inteligencia de datos y la minería de procesos. Tal es el caso de [2] que utiliza data mining en un entorno de Big Data en el correo de China. Debido a la problemática del negocio postal, en dicho artículo se utilizaron técnicas de clustering para agrupar a los clientes según el comportamiento, los hábitos de consumo y el foco de interés logrando generar una estrategia de Marketing postal más precisa y efectiva con resultados muy satisfactorios. Otro caso es [3] donde se aplica process mining en logística buscando similitudes y diferencias entre distintos procesos de entrega en un contexto cambiante de manufactura y logística. En dicho trabajo se comparan diferentes procesos utilizando técnicas de agrupamiento en busca de lograr una documentación automatizada de procesos en un contexto cambiante. En [4] se presenta una metodología que sirve de guía para la ejecución de proyectos de process mining que describe las diferentes etapas. Además se muestra como caso de estudio su aplicación práctica en el proceso de compras de IBM.

En este artículo se utilizarán distintas técnicas de Minería de Procesos para analizar datos basados en la distribución postal de productos en la República Argentina durante los años 2017 a 2019. Los autores de este trabajo desean manifestar que, a la fecha de generación del presente documento, no tienen conocimiento de la existencia de trabajos similares que implementen la minería de procesos en la distribución postal en la República Argentina.

2. Minería de procesos (Process Mining)

El punto de partida de la Minería de Procesos es el registro de eventos. Se asume que el proceso a analizar requiere del registro de una serie de eventos secuenciales pertenecientes a la actividad y que se relacionan con un caso particular. Si bien se pueden almacenar datos adicionales de cada evento deben registrarse obligatoriamente la fecha del evento (día y hora), el identificador del caso y el tipo de evento.

Los tres tipos básicos de minería de procesos [5] son:

- *Descubrimiento*: las técnicas de descubrimiento toman como punto de entrada un log de procesos y generan un modelo sin ninguna información adicional. Un ejemplo es el algoritmo Alpha [6] que toma los datos del log y genera una red de Petri que explica el comportamiento reflejado en el log. Como toda técnica que extrae conocimiento a partir de los datos, la calidad del proceso descubierto dependerá del grado de representación de los eventos relevados con respecto al funcionamiento de dicho proceso. No se pueden descubrirse partes del proceso que no se encuentren representadas por los eventos.
- *Verificación de Conformidad*: Consiste en realizar la comparación de un modelo existente (puede ser el descubierto anteriormente u otro) con la secuencia de eventos real para encontrar desvíos y verificar el funcionamiento. Suelen utilizarse aplicaciones capaces de generar representaciones gráficas y animaciones que posibilitan observar el comportamiento real y ver cuánto se ajusta la realidad al proceso definido originalmente. Su principal ventaja es que muestra la realidad, no es una simulación, por lo que permite realizar un análisis mucho más preciso.
- *Mejoramiento*: Se busca extender o mejorar el proceso existente a través de la información subyacente en la secuencia de eventos. A diferencia de la Verificación de Conformidad, donde se comparan los datos con el modelo, aquí se busca modificar el proceso

Se debe regular el grado de abstracción a utilizar durante el análisis considerando los siguientes aspectos: ajuste, precisión, generalización y simplicidad [1]. Existe una relación de compromiso entre ellos que debe ser tenida en cuenta si se espera obtener buenos resultados. El ajuste se refiere a la habilidad del modelo para explicar el comportamiento observado. La precisión se refiere a la exactitud con la que se ejecuta el proceso; aquí es importante no sobreajustar el modelo a los datos de entrada porque se perderá generalización impidiendo lograr el nivel de abstracción deseado. La simplicidad también se ve comprometida por el sobreajuste ya que se logra agregando mayor detalle a la descripción del proceso.

Finalmente es importante destacar que el resultado obtenido del análisis con minería de procesos está altamente vinculado a la calidad de los datos de entrada. De hecho se sabe que siempre hay una cantidad de ruido en los datos el cual puede deberse a trazas incompletas, intervalos que no se hayan registrado correctamente o duplicidad de datos. Esta información puede distorsionar o falsificar el resultado del análisis [7].

Generalmente es necesario verificar los datos de entrada y realizar un procesamiento previo para eliminar la mayor cantidad de ruido posible.

3. Descubrimiento del proceso de Distribución Postal

En esta sección se llevará a cabo el descubrimiento del proceso a partir de datos basados en la distribución postal de productos en la República Argentina durante los años 2017 a 2019.

El proceso de distribución postal abarca diferentes tipos de productos; en todos los casos el proceso consiste de una recepción del producto del remitente (ya sea por un medio físico o un canal digital) y una entrega al destinatario, pudiendo esta ser exitosa o no. La no entrega del producto no implica un error de proceso ya que puede haber motivos que lo justifiquen.

El proceso registra al menos las siguientes actividades: recepción e identificación, envío para su distribución, uno o más intentos de entrega, espera en centro de distribución y devolución (si no fue posible la entrega). En todos los casos se debe dejar registro de cada uno de los pasos realizados asociados a un identificador único de envío, lo que permite conocer el estado actual del mismo e informarlo al cliente.

Extracción de los datos

El primer paso en el descubrimiento del modelo consiste en la recolección y preprocesamiento de los datos a utilizar. En particular, para este caso de estudio se utilizaron envíos de productos con dos intentos de entrega en el domicilio. El procedimiento actual establece que aquellos productos que no pueden ser entregados se guardan un tiempo a la espera de que el destinatario venga a retirarlo; transcurrido ese tiempo se realiza la devolución al remitente.

Se define una traza como un envío. Cada movimiento que se registre de ese envío es un evento. Se considera una traza completa cuando el envío tiene registrado su ingreso y su fin, con entrega exitosa o no.

Como resultado de la recolección de datos se generó una tabla o muestra de alrededor de 33.000 trazas y un total de 78.000 eventos. Para cada evento se registraron los campos mínimos necesarios para el análisis; estos son: identificador de traza, identificador de evento, fecha de evento, descripción del evento (ver Figura 1). Cada traza puede tener uno o más eventos asociados, se necesita identificador de traza e identificador de cada uno de los eventos para armar la historia de la traza.

De esta forma se dispone de los distintos pasos o eventos que se produjeron a lo largo del envío de cada producto.

Luego se transformaron los datos utilizando el formato XES [8]. XES es una gramática para un lenguaje basado en etiquetas cuyo objetivo es proporcionar a los diseñadores de sistemas de información una metodología unificada y extensible para capturar los comportamientos de los sistemas mediante registros y flujos de eventos [8]. De esta forma se agiliza el manejo de los datos pudiendo ser procesados por distintas herramientas de manera más eficiente.

A fin de facilitar el descubrimiento del proceso correcto, se eliminaron todas las trazas incompletas entendiéndose por tales a aquellas trazas que no tengan o bien un evento de inicio (Recepción) por algún error de carga o bien un evento final (entrega, devolución, destinatario inexistente, fallecido, rechaza el envío o se mudó). Esto último podría ocurrir por error de carga o porque todavía no transcurrió el tiempo estipulado para finalizar el proceso.

Identificador de evento			Descripción del evento	
Identificador de traza				Fecha evento
	123 trazaID	123 EveID	ABC eveDescrip	eveFecha
31	481,053	0	INGRESADO	2017-08-16 10:45:22
32	481,053	2	1 INTENTO DE ENTREGA	2017-08-18 11:15:00
33	481,053	9	DEVOLUCION	2017-08-18 13:00:00
34	481,054	0	INGRESADO	2017-08-16 10:45:28
35	481,054	2	1 INTENTO DE ENTREGA	2017-08-22 12:05:00
36	481,054	9	DEVOLUCION	2017-08-22 17:25:00
37	481,055	0	INGRESADO	2017-08-16 10:45:27
38	481,055	1	ENTREGADO	2017-08-22 15:13:00

Figura 1. Ejemplo de eventos extraídos para el análisis

Utilizando un filtro de reglas heurísticas simples se identificaron las trazas completas dejando sólo aquellas que tienen su estado inicial y los estados finales válidos.

Como resultado de este proceso de filtrado se obtuvieron aproximadamente 16.000 trazas con 43.000 eventos.

3.1. Modelo del proceso

Para generar el modelo del proceso se utilizó un algoritmo clásico de minería de procesos, el algoritmo Alpha, propuesto por primera vez por van der Aalst, Weijters y Märušter[6]. El objetivo de este algoritmo es reconstruir la causalidad a partir de un conjunto de secuencias de eventos. Construye Redes de Petri con propiedades especiales (redes de flujo de trabajo) a partir de registros de eventos (como los que podría recoger un sistema ERP). Cada transición en la red corresponde a una tarea observada.

En lo que se refiere al caso de estudio de la distribución postal, si bien es posible construir el modelo directamente a partir de las 16000 trazas mencionadas previamente, se seleccionaron manualmente algunos casos buscando los más representativos para encontrar el proceso ideal que deben cumplir las trazas, simplificando de esta manera esta etapa de descubrimiento del proceso.

La Figura 2 ilustra la red de petri correspondiente al proceso descubierto que deberán cumplir las trazas.

El proceso encontrado es el siguiente: Todas las trazas deben comenzar con un evento de ingreso y luego sale a distribución. Si se puede entregar, se registra el evento y finaliza el proceso. Si no se puede entregar, en caso de ser un evento final (fallecido, datos faltantes en dirección, no existe dirección, desconocido, etc.), se registra el motivo y finaliza el proceso. Si no es un evento final (por ejemplo se pudo llegar a la dirección pero no había nadie) se registra un primer intento de entrega. Posteriormente se realiza una nueva visita. Si se pudo entregar, se registra el evento y finaliza el proceso. Si en la segunda visita no

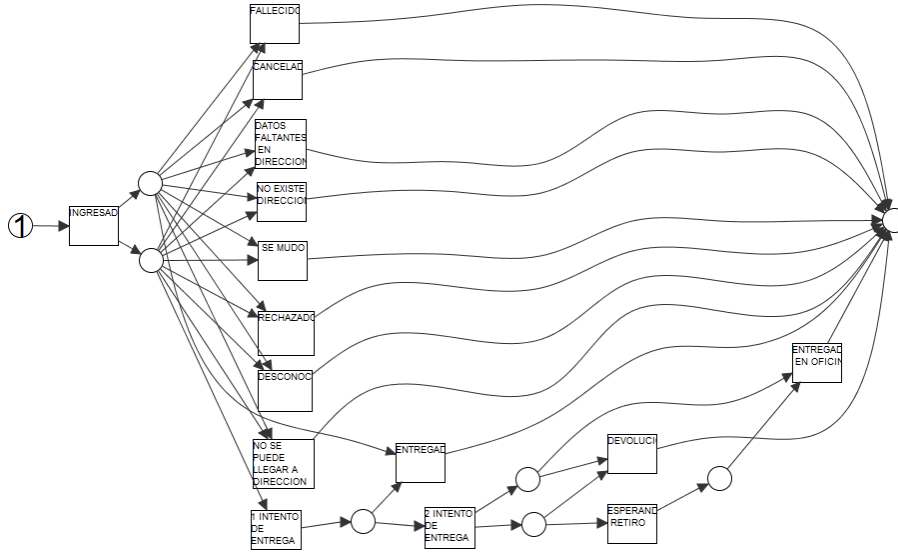


Figura 2. Proceso descubierto que deberán cumplir las trazas

se puede entregar, se guarda el envío en oficina por un tiempo esperando que el destinatario venga a retirarlo y luego se registra la entrega en oficina. Transcurrido el tiempo de espera, si el destinatario no vino a retirar el envío, se registra la devolución. Esos son los posibles escenarios que contempla el proceso.

3.2. Verificación del modelo

El modelo de proceso descubierto en la etapa anterior fue generado a partir de un subconjunto de trazas seleccionadas previamente. En esta sección se llevará a cabo la verificación de conformidad. Tal como se explicó previamente, se buscará determinar cuánto se ajustan las trazas de la muestra al proceso. Se tendrán en cuenta qué pasos se cumplen, cuáles no y si hay algunos adicionales que no están reflejados en el proceso descubierto. Para ello, se comparará cada una de las trazas obtenidas contra el proceso encontrado en el punto anterior. Se utilizará la muestra de datos obtenida en la sección anterior, la cual está conformada por 16.811 trazas con un total de 43.888 eventos.

El resultado se muestra en la Figura 3. Allí se resaltan los eventos más comunes en color oscuro y los recorridos más frecuentes con una línea más gruesa. Se observa que la mayoría de las trazas siguen el camino de la entrega, ya sea en el primer intento o en el segundo.

Las estadísticas de la muestra indican que en un 80 % de los casos la pieza es entregada, ya sea en el primer o segundo intento y que el resto se distribuye uniformemente. En base a estas observaciones se puede afirmar que el 85 % de las trazas se ajusta perfectamente al proceso descubierto con una cantidad promedio de 2.6 eventos en cada traza.

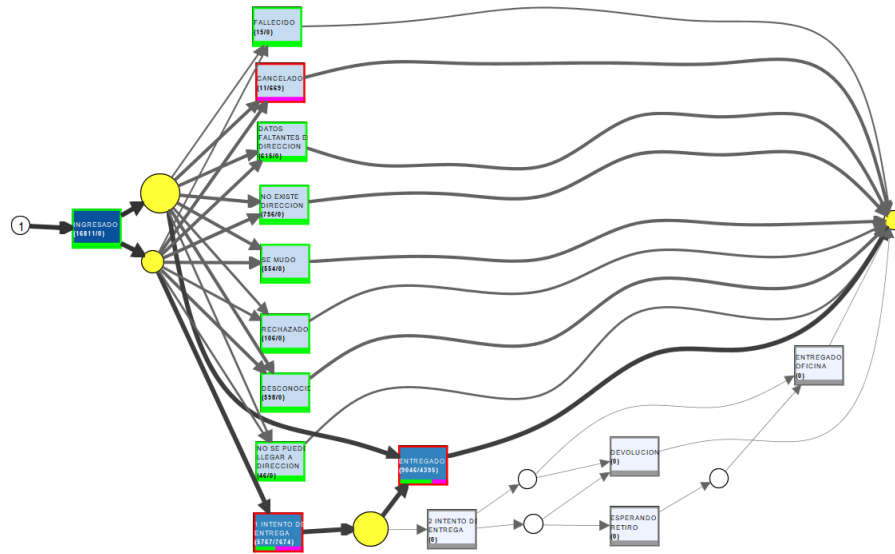


Figura 3. Recorrido de las trazas: eventos más comunes en oscuro y recorridos más frecuentes en línea más gruesa

Sin embargo, algunas trazas no verifican exactamente el proceso descubierto ya sea porque se saltan pasos, o no los hacen en el orden correspondiente o hacen pasos por fuera del modelo. Estas trazas poseerán un valor de ajuste que será más bajo cuanto más se diferencien del modelo. Aquellas que tengan un ajuste menor al 50 % son el objeto de análisis de esta sección ya que representan un gran desvío operativo.

A partir de estos resultados se enfoca el análisis en dos caminos, por un lado las trazas que tienen una excesiva cantidad de movimientos y por otro aquellas que no se ajustan al modelo, ya sea porque hay eventos faltantes o porque no cumplen la secuencia correspondiente.

En el primer caso, dado que la cantidad de movimientos por traza sigue una distribución normal con media 2,6 y desvío 0,95 se consideró adecuado tomar un valor superior a la media mas tres desvíos estándar como un valor representativo de una cantidad excesiva de movimientos.

En base a esto, se procedió a filtrar y posteriormente a analizar las trazas con más de 6 movimientos encontrando 108 casos.

El segundo caso analizado fue el de las trazas que no se ajustaban exactamente al proceso descubierto. En esta oportunidad se fijó el valor del umbral de ajuste en 0.5 por considerar que las trazas que contienen al menos el 50 % de sus movimientos desalineados (no se corresponden) con respecto a lo indicado por el modelo representan una fuerte distorsión del procedimiento y deben ser inspeccionadas. Como se dijo previamente, esto puede ocurrir o bien porque hay

3760173.0 8 events	3780088.0 6 events
INGRESADO #1 02.01.2020 16:29:45.000	INGRESADO #1 20.01.2020 21:00:14.000
1 INTENTO DE ENTREGA #2 13.01.2020 12:25:00.000	1 INTENTO DE ENTREGA #2 21.01.2020 11:12:00.000
2 INTENTO DE ENTREGA #3 14.01.2020 11:20:00.000	1 INTENTO DE ENTREGA #3 22.01.2020 11:16:00.000
2 INTENTO DE ENTREGA #4 15.01.2020 11:20:00.000	ESPERANDO RETIRO #4 22.01.2020 16:04:31.000
1 INTENTO DE ENTREGA #5 16.01.2020 11:20:00.000	DEVOLUCION #5 29.01.2020 15:02:07.000
2 INTENTO DE ENTREGA #6 17.01.2020 09:39:00.000	ENTREGADO #6 31.01.2020 11:15:00.000
2 INTENTO DE ENTREGA #7 20.01.2020 10:10:00.000	
DEVOLUCION #8 24.01.2020 16:45:00.000	

Figura 4. Eventos repetidos e inconsistencias

eventos en la trazas que no aparecen en el modelo o porque los eventos de la traza no cumplen con la secuencia correspondiente.

Como resultado de esto se observó que se trataban de repeticiones de eventos en diferentes días o de registros inconsistentes. La Figura 4 ilustra ambas situaciones. En el cuadro de la izquierda se puede observar un caso en el que aparece registrada una primera visita luego de haber ocurrido previamente una segunda; esta situación sólo puede deberse a un error en el registro. En el cuadro de la derecha de la misma figura se observa una inconsistencia ya que no es posible entregar un producto que previamente fue devuelto.

También es posible utilizar herramientas visuales para generar animaciones que faciliten la comprensión de estas situaciones. En este caso particular, con el conjunto de trazas que tienen un ajuste inferior al 50 % se utilizó Inductive Visual Miner[9] para generar una animación que permite ver cómo se suceden los eventos de cada traza en una línea de tiempo. La Figura 5 ilustra esta animación. En dicha figura cada traza está simbolizada con un token (o círculo) que va recorriendo las diferentes etapas del proceso. Esta representación visual permite observar que unas trazas demoran más que otras y que en ocasiones se presentan algunos comportamientos erróneos. Por ejemplo en la figura 5 se observan movimientos de retroceso en las trazas; en particular, se han marcado con círculos que 25 trazas vuelven del segundo intento al primero y 20 vuelven sobre sí mismas.

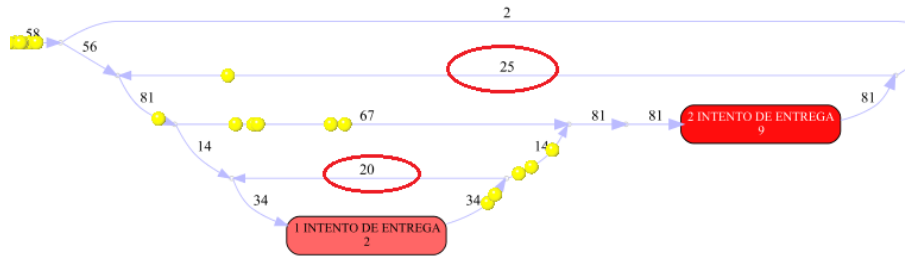


Figura 5. Trazas que no cumplen con el modelo, en círculo la cantidad que retroceden en de un estado en lugar de avanzar (ver sentido de la flecha)

La exportación de las trazas que no se ajustan al modelo permite un análisis detallado de los motivos que ocasionan los desvíos en busca de mejoras al proceso de distribución.

4. Conclusión y futuras líneas de investigación

En este artículo se han utilizado distintas técnicas de Minería de Procesos para analizar datos reales de distribución postal de productos en la República Argentina durante los años 2017 a 2019. No se encontraron a la fecha de generación del presente documento trabajos similares que implementen la minería de procesos en la distribución postal en la República Argentina.

Por medio del modelo generado a partir de trazas seleccionadas se descubrió el proceso que realmente se lleva a cabo. Es importante remarcar el impacto que el criterio de selección de dichas trazas tiene en el modelo obtenido. Los primeros modelos, generados a partir de la totalidad de las trazas, tenían un detalle excesivo de la situación lo cual dificultaba su representación e interpretación posteriores. Actualmente se está trabajando para poder identificar las trazas más frecuentes a partir del modelo inicial para luego filtrar en forma automática las más representativas.

Por su parte, la verificación de conformidad ha permitido determinar situaciones anómalas de interés. Se identificaron tanto los casos que no cumplieron con el modelo como también aquellos que lo hicieron pero fuera de los tiempos esperados o con redundancia de tareas. Se deberá analizar si estos casos representan errores manuales en la carga y buscar una solución.

Se utilizó un umbral de ajuste para establecer el grado de distorsión mínimo que una traza debe cumplir para no afectar el desarrollo del proceso. Este factor debe ser analizado con más detalle para determinar su valor en base al caso de estudio del cual se trate.

Como línea de trabajo futura se propone continuar trabajando con las técnicas de minería de procesos no sólo para modelar situaciones que ya han ocurrido

sino para poder insertar en el sistema alertas tempranas que adviertan de la existencia de desvíos excesivos del modelo.

Referencias

1. Process Mining: Data Science in Action, Wil van der Aalst, 978-3-662-49850-7, 2016, Springer
2. Research of Postal Data mining system based on big data, Xia Hu1; Yanfeng Jin1; Fan Wang, 3rd International Conference on Mechatronics, Robotics and Automation, 2015, 10.2991/icmra-15.2015.124, https://www.researchgate.net/publication/300483008_Research_of_Postal_Data_mining_system_based_on_big_data
3. Context Aware Process Mining in Logistics, Mitchell M. Tseng; Hung-Yin Tsai; Yue Wang, 2017, The 50th CIRP Conference on Manufacturing Systems, <https://www.sciencedirect.com/science/article/pii/S2212827117303311>
4. PM2: a Process Mining Project Methodology, Maikel L. van Eck; Xixi Lu; Sander J.J. Leemans; Wil M.P. van der Aalst, Eindhoven University of Technology, The Netherlands, http://www.processmining.org/_media/blogs/pub2015/pm2_processminingprojectmethodology.pdf
5. Wil van der Aalst, The Process Mining Manifesto by the IEEE Task Force, 2012, <https://www.tf-pm.org/resources/manifesto>
6. Workflow mining: discovering process models from event logs, W. van der Aalst; T. Weijters; L. Maruster, 1041-4347, 2004, IEEE, <https://ieeexplore.ieee.org/document/1316839>
7. Process mining in flexible environments, Christian Walter Gunther, 978-90-386-1964-4, 2009, Technische Universiteit Eindhoven, <https://research.tue.nl/en/publications/process-mining-in-flexible-environments>
8. IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams, IEEE Std 1849-2016, 2016, DOI 10.1109/IEEESTD.2016.7740858
9. inductive visual miner, Sander J.J. Leemans, 2017, <http://leemans.ch/leemansCH/publications/ivm.pdf>
10. Reinventing the Postal Sector in an Electronic Age, Michael A. Crew; Paul R. Kleindorfer, 978-1849803601, 2011, Edward Elgar Publishing