UNIVERSITÀ POLITECNICA DELLE MARCHE

Facoltà di Ingegneria

Laurea Magistrale in Ingegneria informatica e
dell'automazione

Tecniche di process mining applicate a log TSP

Process mining techniques applied to TSP log data

Relatore:
Prof. Claudia Diamantini

Tesi di Laurea di:
Andrea Chiorrini

Anno Accademico 2018/2019

Indice

Sommario

Le transazioni economiche sono sempre maggiormente basate sull'utilizzo di supporti digitali. In particolare i sistemi che si occupano di gestire queste transazioni i Digital Transaction Management System provvedono a accumulare molte informazioni sulle transazioni eseguite. Il presente elaborato si propone di effettuare uno studio sull'applicabilità delle tecniche di process mining al suddetto dominio, confrontando i risultati ottenuti con delle tecniche di data mining classica. I risultati ottenuti mostrano che, sebbene non si possa prescindere dalle tecniche di data mining canoniche come ausilio alle analisi, il process mining consente, attraverso una visione più olistica, di rilevare molte più informazioni nei dati rispetto alle tecniche tradizionali.

1. Introduzione

L'esecuzione di transazioni economiche è sempre più basata sullo scambio di informazioni digitali, che devono pertanto garantire determinati livelli di sicurezza, identificabilità, tracciabilità e non-ripudiabilità. A questo scopo, sono state sviluppate ad esempio tecniche di firma digitale, le quali hanno da alcuni anni in Europa riconosciuta valenza legale, e che sono state quindi integrate in vari sistemi per la gestione e la tracciabilità di transazioni digitali (Digital Transaction Management - DTM). I sistemi DTM producono una quantità significativa di informazioni sullo stato, l'evoluzione, ed eventuali errori delle attività che compongono una transazione che spesso, devono per vincoli normativi, essere mantenute in archivi digitale da poter essere consultati e impugnati in caso di contezioso legale. Il mantenimento di tali informazioni solo per il rispetto delle norme legali si traduce naturalmente in un costo necessario all'esecuzione del business per le aziende che fornisco tali servizi. D'altra parte l'uso delle informazioni prodotte attraverso tecniche avanzate di analisi dati, quali tecniche di machine learning e data mining, potrebbero permettere sia di individuare procedure errate e tentativi di frode, consentendo di sviluppare sistemi più robusti a errori e intrusioni, sia di sviluppare sistemi più snelli e efficienti dal punto di vista dell'utilizzatore così da migliorare il servizio fornito e potenzialmente ottenere dei profitti superiori per l'azienda attraverso il vantaggio tecnologico nei confronti dei suoi competitors.

La natura di tali informazioni risulta però molto spesso ostica sia alla lettura da parte di un essere umano sia all'elaborazione e all'identificazione dei contenuti rilevanti da parte di un calcolatore.

Il presente elaborato si propone di avviare una prima analisi esplorativa dei log forniti dalla società © Namirial S.p.A attraverso tecniche di data mining e di process mining per valutare le potenzialità di tali approcci in questo dominio applicativo.

In particolare è stata prima effettuata un'analisi descrittiva per individuare situazioni potenzialmente rilevanti dal punto di vista del business, poi si sono analizzate tale situazioni utilizzando una "prospettiva a processi", cioè utilizzando tecniche di process mining per estrarre modelli del processo tipico di firma.

I risultati ottenuti dimostrano come la causa principe di ritardi nella firma tempestiva da parte di un utente sia la non sottoscrizione alla prima apertura del documento: quasi tutti i casi "lenti" sono svolti in più riaperture del documento. Inoltre, è stato rilevato un impatto considerevole sulle firme apposte dopo più aperture del caso in cui la firma vada effettivamente disegnata attraverso l'interfaccia digitale per essere apposta. Infine in relazione ai risultati ottenuti è stato possibile stabilire che il process mining trova un'applicazione anche in questo dominio, dal momento che la visione più olistica che riesce a garantire all'analista permette di rilevare informazioni che sfuggirebbero utilizzando solo delle tecniche "classiche" di data mining.

Il presente articolo si sviluppa pertanto come segue.

Nel primo capitolo si fornisce un excursus sulla metodologia adottata, la quale farà anche da scheletro per la struttura dell'elaborato. Nel secondo capitolo si illustrerà il dominio in cui si è operato, con riferimento alle prime tre fasi della metodologia utilizzata. In seguito nel terzo capitolo si mostreranno le implementazioni utilizzate e i si commenteranno i risultati ottenuti.

Infine seguirà una conclusione derivata dalla analisi svolte e una descrizioni di possibili direzioni di futuri lavori.

2. Metodologia

L'approccio utilizzato nel presente elaborato è basato sulla metodologia CRISP-DM (CRoss-Industry Standard Process for Data Mining) [1]. Questa metodologia è un open standard e risulta essere uno dei modelli maggiormente utilizzati nell'ambito della data analytics.

CRISP suddivide il processo di data mining e più in generale di data analytics in sei principali fasi:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Nessuna delle fasi è rigorosamente bloccata, anzi il processo stesso è concepito come altamente flessibile risultando possibile muoversi liberamente tra le varie fasi in base ai risultati, anche parziali, di uno step: questo fenomeno risulta naturale dal momento che quando si scoprono delle nuove informazioni come risultato di una fase esse possono generare ulteriori curiosità o dubbi su informazioni acquisite precedentemente, possibilmente, anche per comprendere al meglio le informazioni

appena ottenute. In figura 1.1 riportiamo un diagramma che evidenzia con delle frecce le transizioni più comuni che occorrono nel flusso di lavoro utilizzando CRISP-DM.

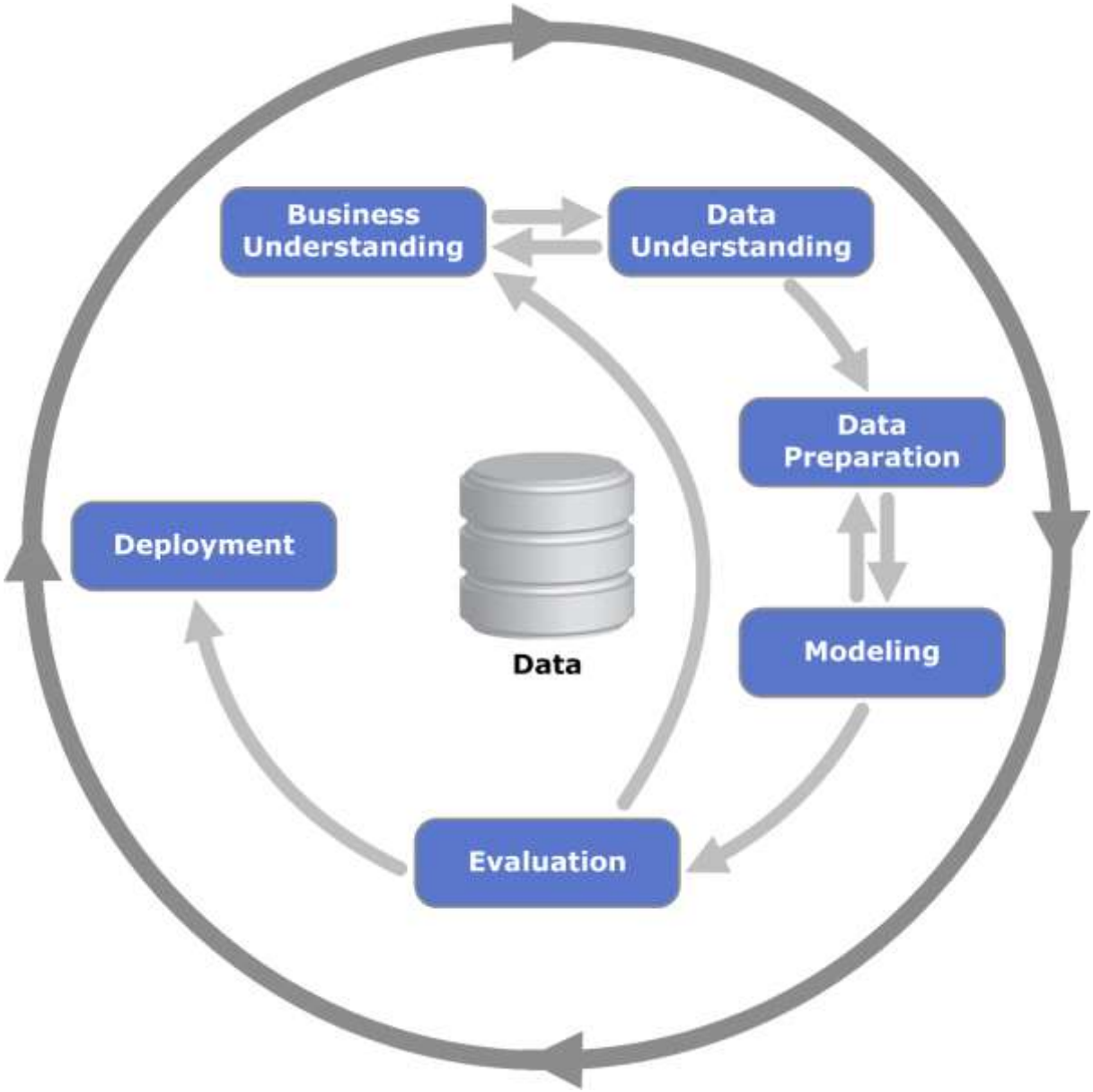


Fig. 1.1

3. Comprensione del dominio

3.1 Business Understanding

La comprensione del metodo di business che stiamo analizzando è passata inizialmente attraverso due differenti interviste con esperti di dominio; dopodiché a ogni dubbio o domanda che l'ulteriore conoscenza dei dati portava a sorgere si è sempre provveduto ad avere un confronto, a tal riguardo, con almeno uno di questi. Si è altresì provveduto ad avere un'attivazione di un account di test con tutte le funzionalità del prodotto studiato attive, così da poter effettuare test di prima mano su tutti i possibili use case che un utente firmatario poteva trovarsi innanzi.

In particolare il prodotto analizzato è eSignAnyWhere una portale web per la firma digitale conforme alle normative europee vigenti. Proprietà di Namirial S.p.A, il software si compone di un'interfaccia utilizzabile dal cliente della società per arricchire il documento pdf, che vuol far sottoscrivere, con dei campi che possono essere compilati dall'utente finale, e in particolare con alcuni di questi campi deputati a poter essere firmati attraverso vari metodi possibili con un certo grado di valenza legale in relazione alla procedura utilizzata per far apporre la firma.

Senza scendere troppo nel dettaglio normativo l'aspetto maggiormente rilevante da rimarcare è come vi siano vari tipi di firma utilizzabili in relazione alla forma contrattuale sottoscritta con Namirial, ma che la forza legale delle firme apposte

derivi principalmente dalla procedura utilizzata per far apporre la firma e meno dalla forma con cui questa viene apposta: ad esempio se sia sufficiente un click per firmare o se sia necessario disegnare la propria firma. Naturalmente a questo fanno eccezione i casi in cui la forma è dichiaratamente parte integrante della procedura prevista per quel grado di valenza legale della firma.

Altresì rilevante è come ogni firma sia di fatto un hash associato al documento sottoscritto e che il mantenimento agli atti, ovvero in un qualche archivio, di tale documento sia una parte fondamentale per garantire l'impugnabilità della firma apposta in sede legale. eSAW provvede solamente alla costituzione del documento sottoscritto, il servizio di archiviazione può essere acquistato separatamente sempre da Namirial dagli interessati oppure può essere svolto indipendentemente.

Il corretto mantenimento della documentazione associata alla sottoscrizione di un contratto non prevede esclusivamente il contratto sottoscritto in sé e l'hash associato alle firme, ma anche un ulteriore documento di log chiamato audittrail che riporta tutto il processo svolto che ha portato alla firma del documento in questione.

Tali audittrail sono dei file XML che vengono generati automaticamente dal sistema alla chiusura di una pratica. La parte maggiormente rilevante per i scopi del presente elaborato di tali files è rappresentata dalla porzione che traccia l'interazione dell'utente con il documento.

Un ulteriore aspetto interessante è che l'applicazione lavoro con il concetto di pratica e una pratica può avere più firmatari che devono sottoscrivere il documento magari in punti diversi o in momenti successivi, ma sebbene questa ipotesi sia contemplata dal software quasi sempre l'applicativo viene utilizzato avendo un unico destinatario firmante per ogni pratica.

3.2 Data Understanding

Anche in questa fase l'uso dell'account di test di eSAW è risultato fondamentale insieme a svariate prove per comprendere la semantica di molti dati presenti negli audittrail. Si sono visualizzati e analizzati svariati files XML relativi sia a audittrail reali sia a audittrail di test. Per coadiuvare questo processo si sono prodotti sia dei script python sia delle dashboard interattive sviluppate attraverso Qlik, in fase di modeling, che producessero delle statistiche preventive sul dataset: in questo modo è risultato possibile comprendere più agevolmente quali dei vari XML fosse più utili visualizzare direttamente.

Procediamo ora a illustrare la conoscenza ottenuta del dataset e derivata da varie iterazioni di questa fase.

Il dataset che abbiamo analizzato è composto da 29347 audittrail relativi a 7 differenti organizzazioni. Ciascun audittrail poi essendo un file XML ha una struttura

a albero che sebbene variabile mantiene alcuni capisaldi strutturali che riportiamo di seguito.

Sotto il nodo radici “AuditTrail” che possiedi un attributo relativo alla versione di eSAW utilizzata e uno alla data di creazione di XML abbiamo i seguenti nodi figlio:

- EnvelopeId
- EnvelopeName
- EnvelopeStatus
- EnvelopeCreationDate
- EnvelopeSendDate
- EnvelopeExpirationDate
- Sender
- ElectronicDisclosures
- Recipients
- Notifications
- SendFinishedDocuments
- PreventMailSending
- AttachSignedDocuments
- Signature

Come visibile i primi nodi sono relativi al concetto di pratica e non hanno figli nonché inoltre specificano informazioni di servizio interne sulla pratica specifica:

questi nodi non sono rilevanti per le analisi effettuate poiché le informazioni che possono interessarci sono anche riportate in altri porzioni dell'albero.

Il concetto di "Sender" neanche risulta interessante per le nostre analisi visto che gli auditrail che abbiamo a disposizione sono relativi a tutte compagnie che hanno un unico "Sender" per tutta l'organizzazione.

Tutti gli altri nodi fanno riferimento a ulteriori informazioni di servizio, per garantire, ad esempio, la conformità a determinate procedure opzionali per avere gradi più alti di valenza legale: questi non sono quindi pertinenti con gli scopi di questo elaborato ad eccezione di Recipients dentro il quale vengono racchiusi i vari destinatari della pratica e tutte le informazioni relative loro, nonché alcune informazioni ridondate riguardo la fase della pratica relativa a quello specifico destinatario. Nella quasi totalità dei casi nel nostro dataset vi erano solo pratiche con un unico destinatario. Di seguito si riporta anche l'elenco dei nodi figlio di un qualunque nodo di tipo Recipient:

- FirstName
- SealingProfileName
- LastName
- Type
- FinishDate
- Status

- RejectReason
- WorkstepId
- History
- AuthenticationMethods
- MailSubject
- MailContent
- WorkStepInformation
- auditTrail
- PreventMailSending

Per amor di brevità e perché esula dagli scopi dell'elaborato si risparmia la spiegazione di ciascun nodo e si rimarca solo come alcuni siano opzionali, alcuni spesso inutilizzati o non particolarmente informativi e altri ancora ridondanti. Il nodo che maggiormente ci interessa è il nodo "auditTrail" e si noti che l'iniziale è minuscola a differenza del nodo radice dell'albero. L'attenzione viene rivolta principalmente a questo nodo perché è il nodo che al suo interno mantiene la maggior parte del contenuto informativo di tutto il documento e mantiene anche delle informazioni presenti in altri nodi.

In particolare "auditTrail" risulta rilevante per i nostri scopi perché ha come figli gli eventi occorsi durante il processo di firma, pertanto può avere un numero di figli

totalmente variabile, abbiamo infatti casi in cui il numero di eventi è solo 5 come anche casi in cui il numero di eventi è oltre 400.

Gli eventi fondamentalmente sono suddivisi in 3 famiglie principali:

1. CreateSignAnywhereAuditTrailEvent
2. BasicAuditTrailEvent
3. ExternalAuditTrailEvent

La classe di eventi 1 rappresenta, come si evince dal nome, l'evento relativo alla creazione di un audittrail è pertanto sempre presente. Più nel dettaglio è relativo alla creazione della richiesta di una firma, o di un atto di riconoscimento, di un documento, a un destinatario. Si tratta di una classe di eventi che al loro interno hanno un unico tipo di evento chiamato "WorkstepCreated" il cui concetto è appunto quello suddetto e che risulta sempre essere il primo evento di ogni audittrail.

La classe di eventi 2 rappresenta quegli eventi relativi specificatamente alla procedura di firma o comunque che hanno una valenza sulla burocrazia della pratica. In questa categoria ricadono:

- PrepareAuthenticationSuccess: rappresenta il riconoscimento da parte del sistema del successo nella preparazione della procedura di autenticazione dell'utente destinatario della pratica.

- SignWorkstepDocument: rappresenta il riconoscimento dell'apposizione con successo di una firma in un campo del tipo appropriato.
- WorkstepFinished: rappresenta il successo della sottoscrizione della pratica nella sua totalità da parte del destinatario firmatario corrente; non necessariamente coincide con la fine di tutte le attività dell'utente rispetto al documento e quindi con l'ultimo evento nell'audittrail.
- WorkstepRejected: rappresenta il rifiuto della sottoscrizione del documento inviato, e quindi il suo fallimento, anche questo non coincide necessariamente con l'ultimo evento in un audittrail.
- FinishWorkstepOnPageLoad: è presente solo in pratiche inviate con il solo scopo di presa visione certificata del destinatario, rappresenta il successo della recapita dal momento dell'apertura del documento per la prima volta.
- AuthenticationSuccess: rappresenta il successo da parte dell'utente nell'autenticarsi correttamento verso il sistema.
- AuthenticationFailed: rappresenta il fallimento da parte dell'utente nell'autenticarsi correttamento verso il sistema.
- AddedAttachment: rappresenta il successo nell'operazione di inserimento di un allegato alla pratica da parte del firmatario.
- PreparePayloadForBatch: rappresenta il riconoscimento da parte del sistema del successo nella preparazione della procedura di firma in blocco di vari campi firma presenti nel documento.

- StartBatch: rappresenta l'avvio della procedura di firma in blocco da parte dell'utente.
- EndBatch: rappresenta la conclusione con successo della procedura di firma in blocco.
- UndoAction: rappresenta l'annullamento dell'ultima azione effettuata dall'utente da parte dell'utente stesso.
- DocumentDownloaded: rappresenta il download del documento da firmare da parte dell'utente, nello stato corrente, quindi anche parzialmente compilato.
- FlattenedDocumentDownloaded: come "DocumentDownloaded" ma compresso.
- FormsFilled: rappresenta l'atto di compilazione di un campo form da parte del destinatario
- AuditTrailRequested: rappresenta l'atto di richiesta e il download del documento audittrail da parte del destinatario in formato pdf, questa azione forza il sistema a produrre un audittrail riportante tutto ciò che è successo fino a quel momento, anche se la pratica non è conclusa.

La classe di eventi 3 invece è quella relativa all'interazione dell'utente con il documento da firmare escluse però quelle operazione che ottengono una valenza per

la tipologia 2. Di conseguenza gli eventi che ricadono in questa categoria sono i seguenti:

- PageViewChanged: rappresenta il cambio di pagina visualizzata, avviene anche quando il documento viene aperto da capo.
- CalledPage: rappresenta l'atto da parte dell'utente di aprire il documento per la prima volta o a seguito di una chiusura dello stesso.
- WhoIsInformation: rappresenta l'atto di acquisizioni delle informazioni sull'utente che il sistema è autorizzato dall'utente ad acquisire, perlopiù di natura geografico-spaziale
- Draw2SignDialogClosed: rappresenta la chiusura del pop-up di firma di un campo senza aver effettivamente apposto o confermato la firma, per le firme che vanno disegnate.
- Click2SignDialogClosed: rappresenta la chiusura del pop-up di firma di un campo senza aver effettivamente apposto o confermato la firma, per le firme che vanno apposte solamente cliccando i pulsanti di conferma.

Ognuno di questi eventi poi ha al suo interno sempre un timestamp che fissa il momento in cui sono occorsi. Sono poi presenti ulteriori informazioni in base al tipo del nodo, e una serie di nodi figli per ospitarle, se necessario.

Sebbene sia parte della prima iterazione della fase di modeling si riporta in figura 3.1 la numerosità degli eventi su tutto il dataset suddivisa per tipo di evento.

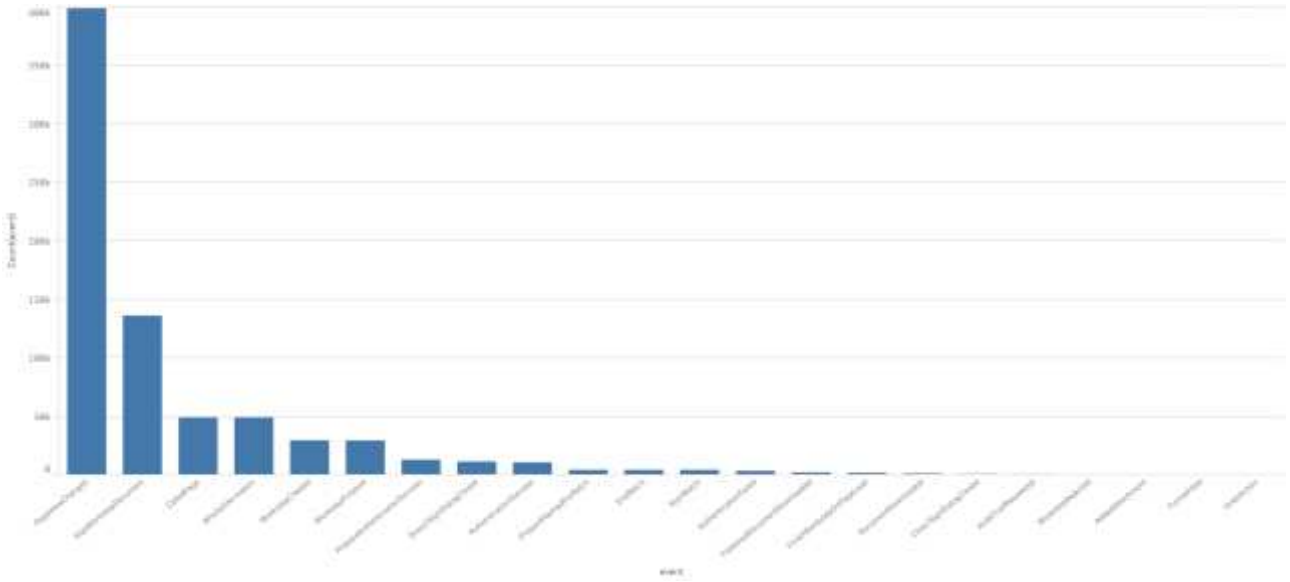


Fig. 3.1

Si può facilmente notare come alcuni degli eventi presentati rappresentino una trascurabile minoranza di occorrenze, e non tanto rispetto a “PageViewChanged” che fa risultare trascurabile qualunque tipo di evento ma rispetto anche a tutti gli altri come mostrato in figura 3.2 rimuovendo appunto “PageViewChanged” e “SignWorkstepDocument”.

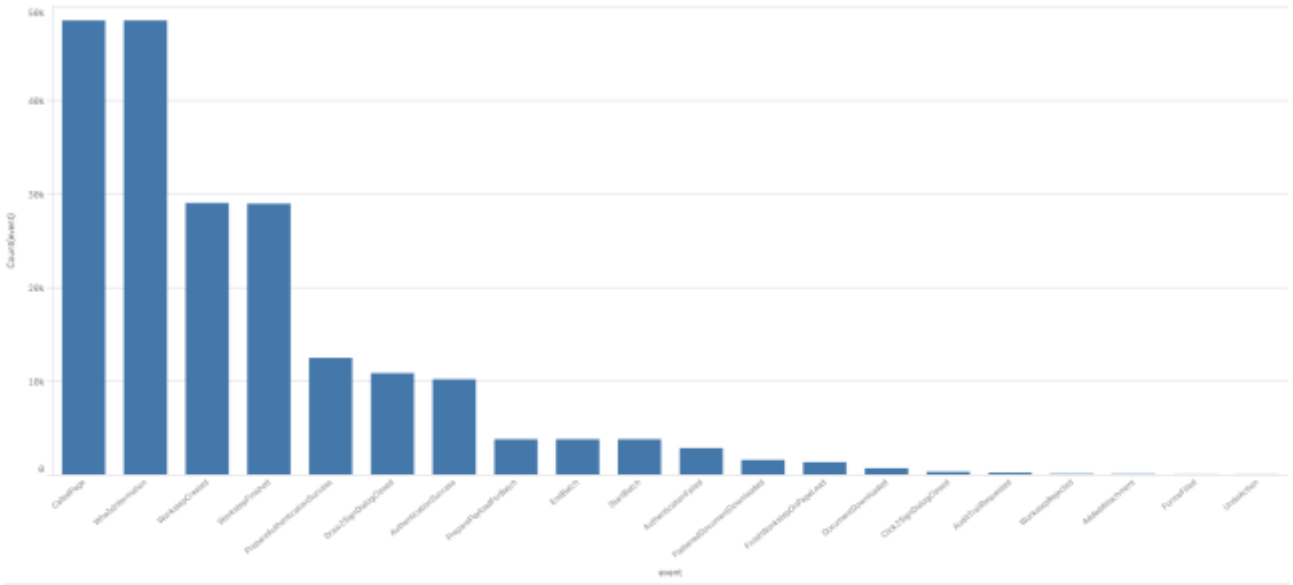


Fig. 3.2

3.3 Data Preparation

La fase di data preparation del nostro progetto ha di fatto mappato fedelmente una procedura di ETL piuttosto standard: almeno nella prima iterazione. In primo luogo i dati non erano disponibili in un sito di storage a cui avessimo l'accesso fisico per scaricarli agevolmente in modo massivo e pertanto è stato necessario implementare una procedura di estrazione automatica attraverso chiamate SOAP: le uniche supportate da eSAW per le interrogazioni. I dati richiesti venivano scaricati codificati in base64 e quindi la necessità iniziale era di decodificarli così da avere degli ASCII leggibile. Fatto ciò, dopo aver eseguito un'iterazione di data understanding sui dati reali si sono avviati dei primi script di estrazioni di dati e metadati sugli audittrail a

disposizione. Abbiamo ricavato dati relativi alle pratiche e ai destinatari, come ad esempio: il numero di firme da apporre e form da compilare per documento, o il numero di firmatari, destinatari in copia e destinatari in presa visione con le informazioni geospaziali della connessione e l'operatore del servizio internet con cui erano connessi o anche informazioni relative ai vari tempi che avevano caratterizzato ogni pratica.

All'iterazione successiva si era mostrato necessario estrarre dati complessivi sulla numerosità e tipologia degli eventi occorsi per pratica nonché sulla durata di questi eventi: intendiamo per durata il tempo intercorso tra l'occorrenza di un evento e il suo successivo.

Da qui poi si è rivelato necessario effettuare una pulizia del dataset dai dati relativi a pratiche con più di un destinatario firmatario perché erano troppo pochi per essere utilizzati e allo stesso tempo inconsistenti con lo scenario più comune: quello delle pratiche con firmatario unico.

Un altro problema riscontrato, era relativo all'inconsistenza di alcuni timestamp degli eventi nel dataset, i quali causavano una durata temporale di alcuni eventi negativa. La questione è stata approfondita anche con dei controlli manuali che hanno verificato un'inconsistenza logica su un'ulteriore minoranza del dataset la quale è stata anch'essa rimossa in questa fase di pulizia.

Il risultato derivato da questa fase su cui si è maggiormente lavorato è stato la trasformazione dell'elenco di eventi tipico degli eventi degli audittrail in un formato che fosse conforme alla nozione di traccia [2] così da poterlo manipolare con degli strumenti deputati al process mining e delle manipolazioni di preprocessing ulteriore attraverso questi.

Infine volendo porre l'accento esclusivamente sulla durata delle azioni dell'utente si sono volute collassare le pause temporali in cui il documento era chiuso tra la prima apertura e il completamento del processo così da poter individuare le eventuali azioni che effettivamente rallentassero il processo. Pertanto si è provveduto a modificare i vari timestamp delle azioni in modo conforme a tale obiettivo.

4. Analisi del dominio

4.1 Modeling

La fase di modeling della presente campagna di analisi ha richiesto inizialmente la costruzione di alcune dashboard interattive che ci consentissero di navigare i dati per una primissima fase esplorativa: a questo scopo per l'implementazione si è utilizzato il software Qlik Sense. Individuati i primi elementi rilevanti su cui focalizzare l'analisi si sono costruite ulteriore dashboard che si specializzassero nella visualizzazione degli aspetti reputati rilevanti.

A questo punto si è avviata la costituzione di modelli derivati attraverso il process mining sia attraverso il tool ProM sia attraverso Disco.
Su ProM si è utilizzato il Data-aware Heuristic interactive Miner, il quale permette una rapida esplorazione interattiva dello spazio dei parametri utilizzando varie euristiche. Questo strumento utilizza gli attributi dei dati per migliorare la procedura di discovery del processo e fornire un servizio di conformance checking built-in [3].

In seguito per avere maggiore sicurezza sui modelli ottenuti abbiamo anche utilizzato Disco così da avere il risultato derivato da un secondo algoritmo e allo stesso tempo una visualizzazione più piacevole dal punto di vista grafico. Disco per generare i suoi modelli utilizza il fuzzy miner [4], un algoritmo il cui approccio si focalizza

principalmente nel poter mantenere diversi gradi di fedele semplificazione e astrazione del processo che si intende generare.

In particolare la nostra analisi attraverso tali strumenti si è orientata alla ricerca e evidenziazione del ruolo di eventuali azioni dell'utente che caratterizzassero la differenza tra i processi considerati brevi e quelli considerati lunghi.

4.2 Evaluation

La fase di evaluation che di fatto corrisponde all'analisi vera e propria è anche dove presentiamo i risultati effettivi che abbiamo ottenuto spiegando il flusso che ha portato alle varie iterazione delle altre fasi. Dalla implementazione delle dashboard interattive per una prima fase esplorativa del dataset ci si è mossi per di rilevare la presenza di eventuali caratteristiche anomale di una porzione dei dati.

Di seguito si riportano i diagrammi che hanno permesso di rilevare le successive direzioni della campagna di analisi.

Il foglio che ha evidenziato maggiormente la presenza di problematiche viene riportato in figura 4.1 e 4.2: in particolare, si può notare come il tempo medio di firma di un documento, dopo l'apertura dello stesso da parte dell'utente, sia di 4.42 ore, ovvero 4 ore e 25 minuti circa mentre limitando i dati a quelli relativi alle pratiche firmate in meno di un giorno il tempo scenda a 0.78 ore ovvero circa 45 minuti (in Fig. 4.2).

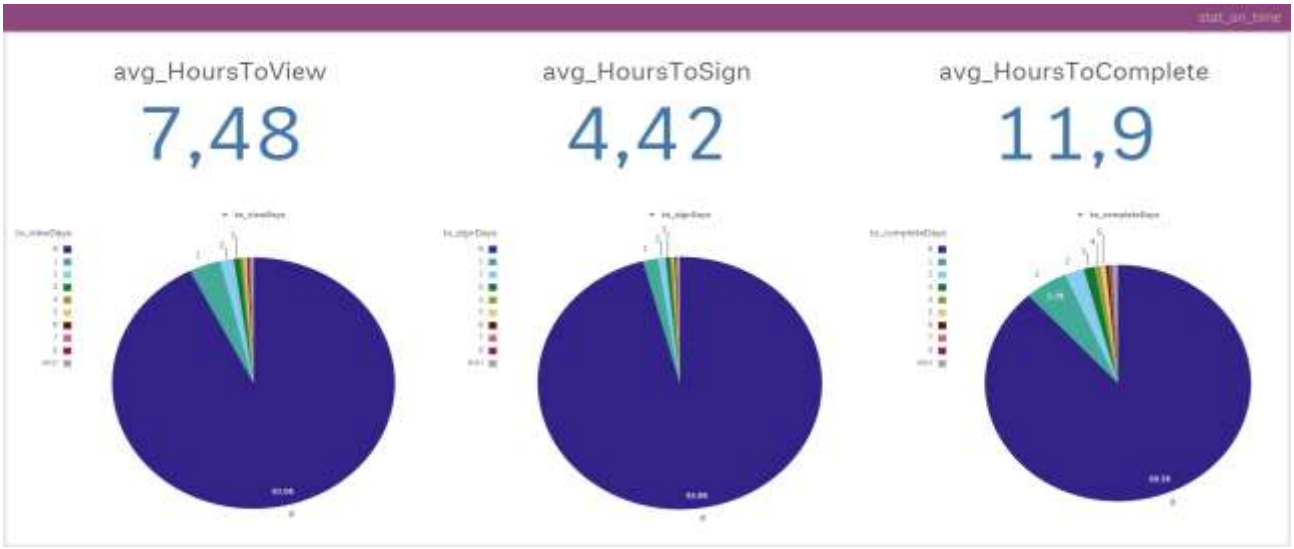

avg_HoursToView
7,48
avg_HoursToSign
4,42
avg_HoursToComplete
11,9

Fig. 4.1

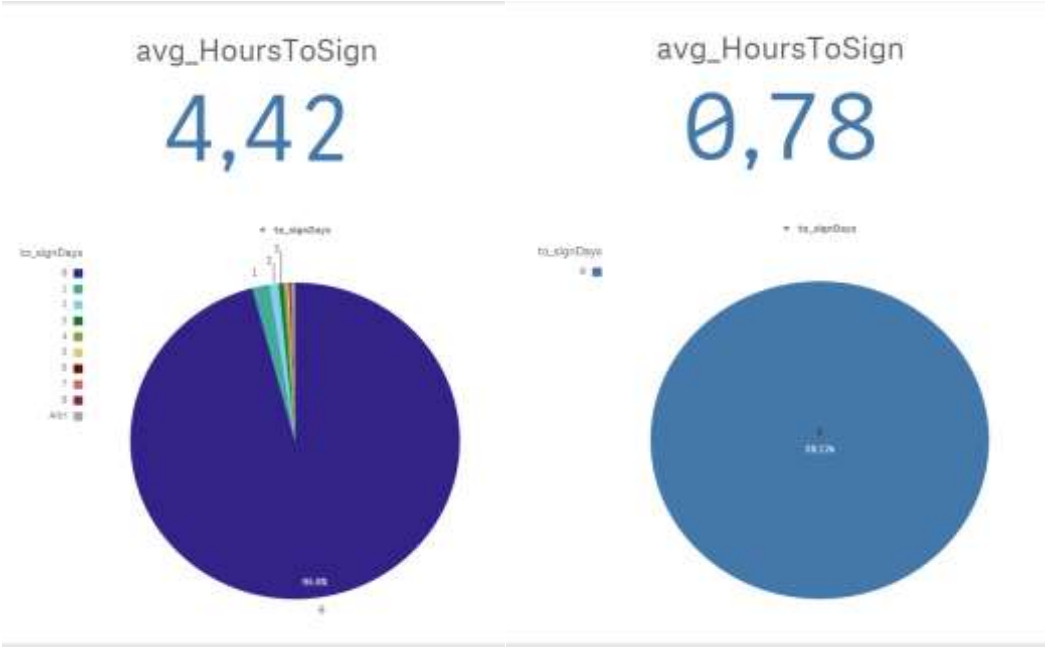

avg_HoursToSign
4,42
avg_HoursToSign
0,78

Fig.4.2

A partire dall'informazione relativa a questa grande escursione nel tempo di firma delle pratiche si è ritenuto opportuno procedere alla costruzione di un istogramma che rappresentasse la distribuzione delle pratiche in relazione al tempo di firma delle stesse, figura 4.3.

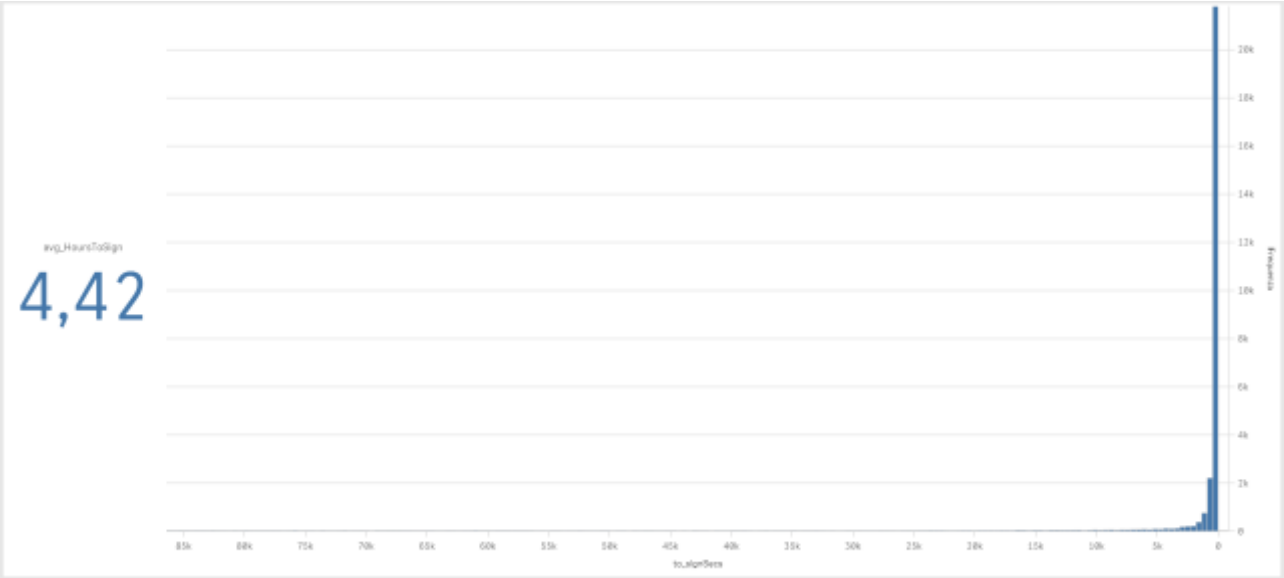


Fig. 4.3

Si nota facilmente come l'andamento dei tempi sia particolarmente sbilanciato, per approfondire l'entità di questo sbilanciamento abbiamo provveduto a effettuare dei filtri successivi per determinare l'impatto che le varie porzioni del dataset hanno sul tempo medio di firma, di seguito si riportano i grafici relativi alla soglia più significativa rintracciata: in figura 4.4 selezionando le prime due barre, ovvero l'80% delle pratiche, mentre in figura 4.5 si sono selezionate le restanti successive.

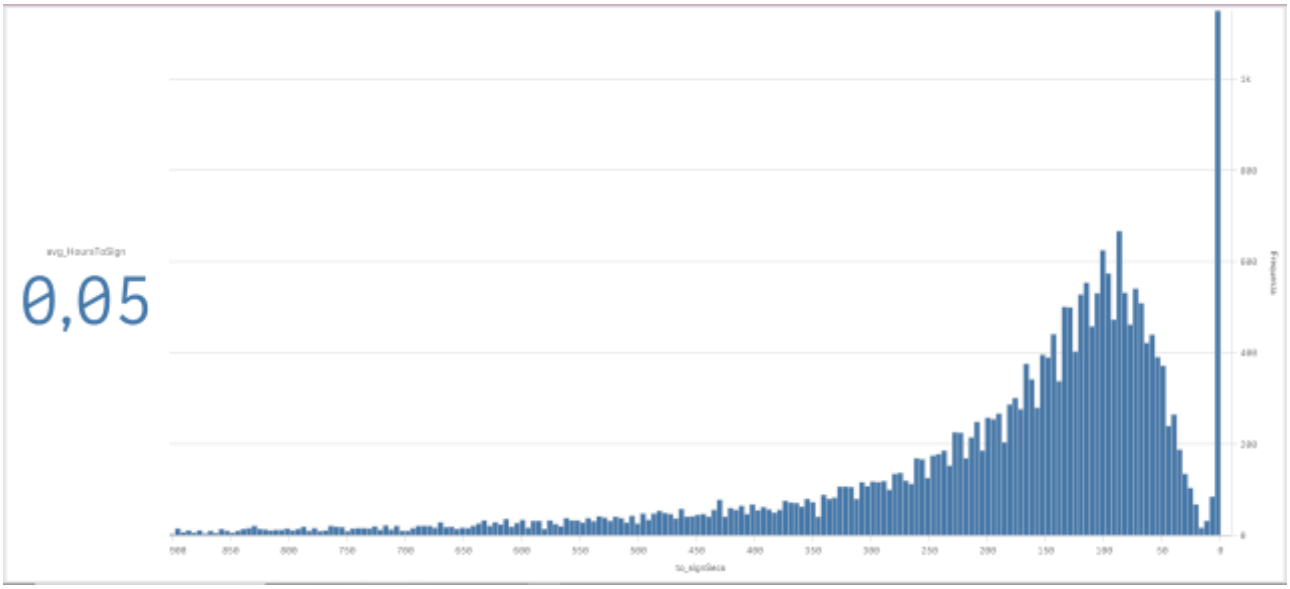


Fig. 4.4

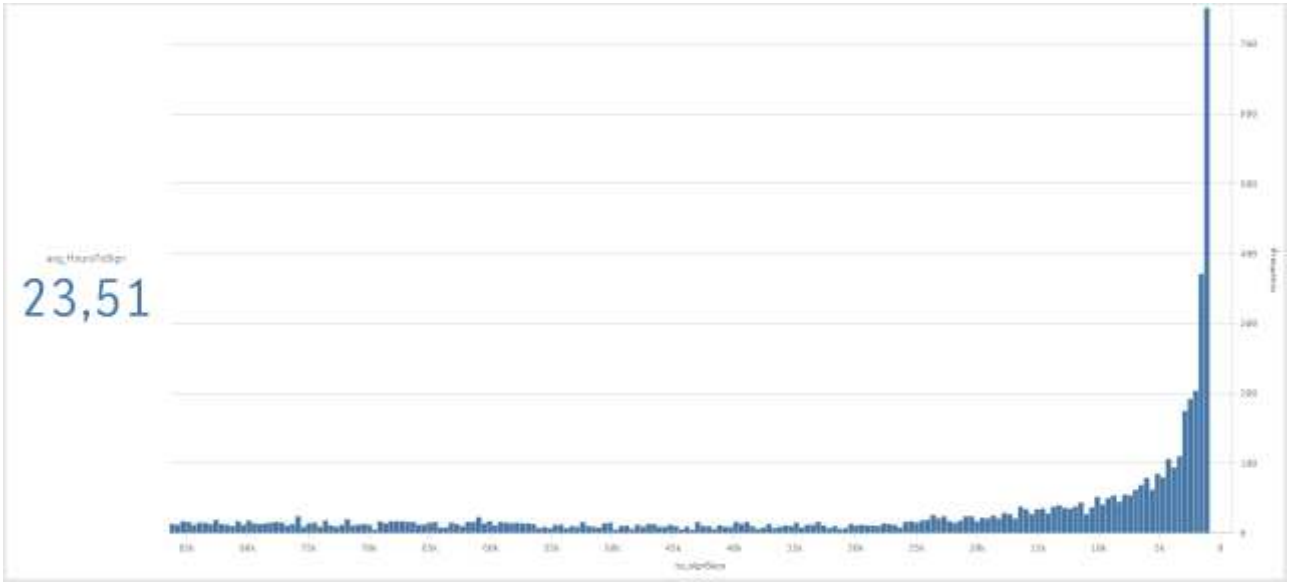


Fig. 4.5

Dalle due immagini si nota facilmente anche controllando il KPI sulla sinistra come il contributo principale alla crescita del tempo medio sia dovuto a solo il 20% delle pratiche.

Notando questo fatto abbiamo cercato di specializzare ulteriormente la ricerca andando a focalizzarci sui singoli eventi che compongono i vari audittrail e le loro durate, cercando opportune caratterizzazioni per gli auditrail di breve durata, inferiore ai 15 minuti, e per quelli di lunga durata, superiore ai 15 minuti.

Come primo aspetto si è esplorata l'effettiva composizione solo in termini di eventi occorsi durante i due tipi di audittrail.

In figura 4.6 riportiamo le occorrenze di eventi presenti nella totalità degli audittrail di breve durata, mentre in 4.7 l'analogo per gli audittrail a lunga durata.

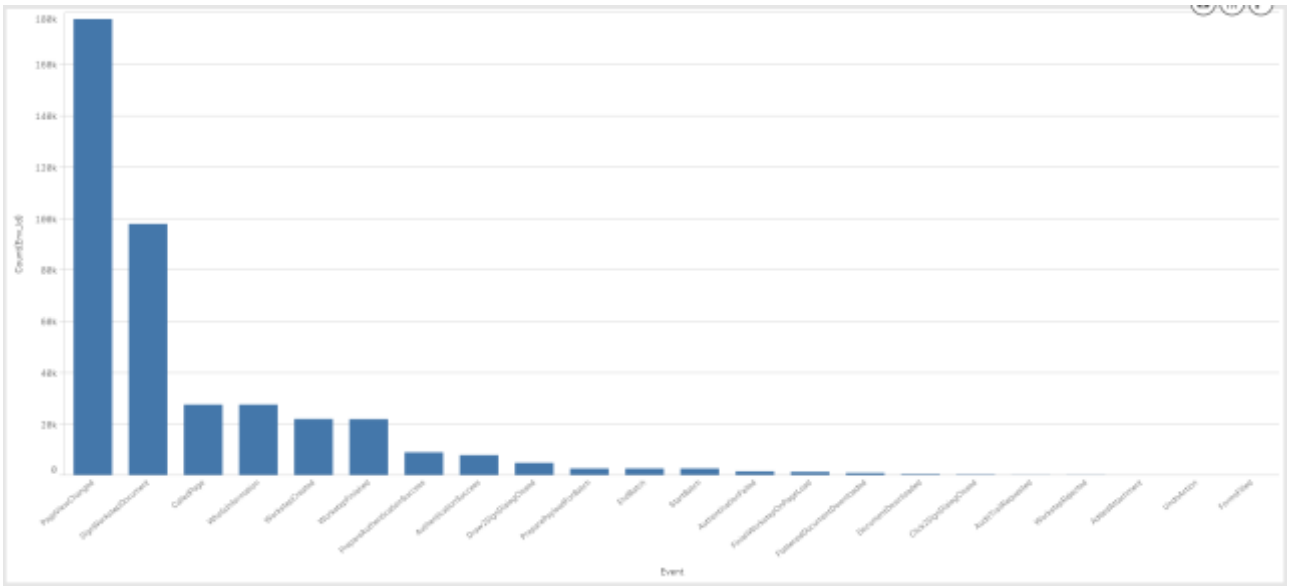


Fig. 4.6

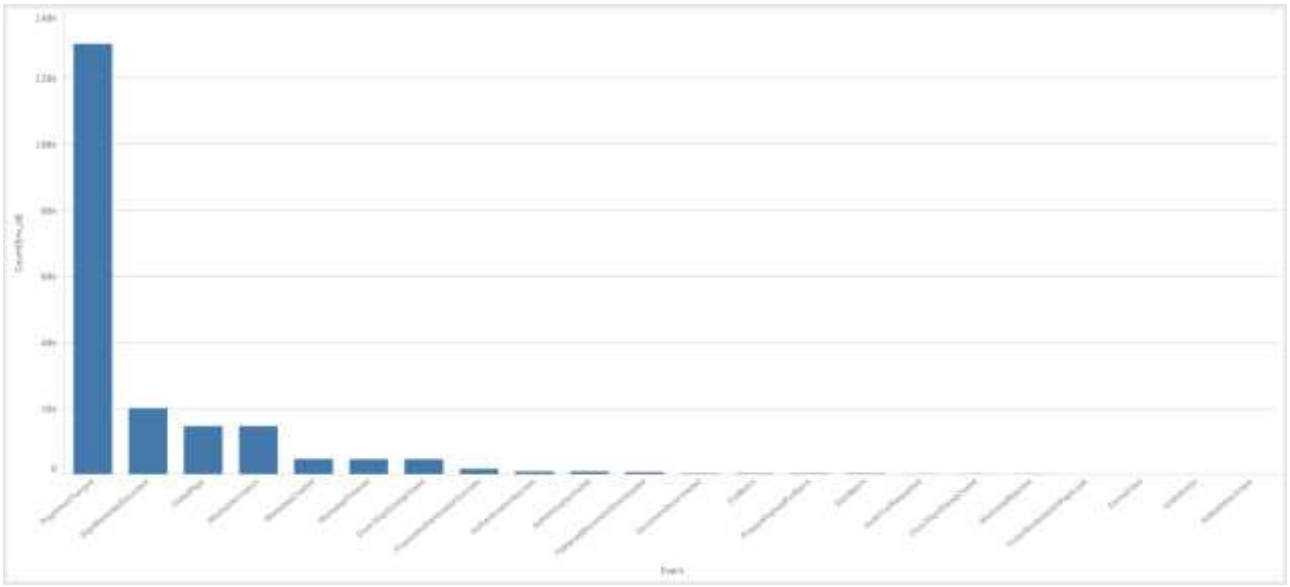


Fig. 4.7

Innanzitutto, per valutare i grafici si ricorda come l'evento "WorkstepCreated" rappresenti la creazione di una pratica e, di conseguenza, il numero di pratiche che ricadono in ciascuna delle due categorie. Ulteriore fatto da ricordare è che l'evento "PageViewChanged" rappresenta la visualizzazione da parte del firmatario di una nuova pagina, perciò la numerosità molto elevata di questo evento risulta essere in linea con ciò che ci si aspettava risultando il più comune in entrambi i grafici. Interessante inoltre è la numerosità relativizzata di questo evento rispetto al numero di pratiche, essa è molto più alta arrivando a un rapporto di circa 26:1 nel caso delle firme a lunga durata contro il 9:1 di quelle a breve durata.

Un'altra differenza peculiare tra i due grafici è relativa alla posizione in cui è collocata la classe di eventi "Draw2SignDialogClosed": questa categoria è alla

settima posizione nel diagramma relativo alle firme lunghe mentre è alla nona in quello relativo alle brevi.

Altro fatto molto interessante è relativo alla categoria “CalledPage”, rappresentante l’apertura da parte dell’utente del documento da firmare. Questo tipo di evento nelle pratiche lunghe è presente con un rapporto di quasi 4:1 rispetto al numero di pratiche, mentre nel caso breve il rapporto supera di poco 1:1.

Ciascuno di queste informazioni presa singolarmente ha una certa rilevanza ma il loro valore salirebbe drasticamente se venissero contestualizzate. In particolare la maggiore frequenza di alcuni tipi di eventi, seppur è di per sé inutile, ci ha indotto a ricercare intorno a questi le principali ragioni per cui alcune firme richiedono drasticamente più tempo attraverso l’uso di strumenti di process mining utili per, appunto, ottenere informazioni circa il flusso degli eventi che compongono le singole istanze del log di firma.

Dopo aver trasformato il dataset portandolo a un formato coerente con la nozione di “traccia” abbiamo avviato un’analisi esplorativa di carattere più olistico attraverso l’heuristic interactive miner di ProM.
La possibilità di avere l’interattività per una variazione dinamica dei parametri dell’algoritmo è risultata essere fondamentale in questa fase, dal momento che sebbene avessimo delle direzioni di ricerca non era chiaro quanto fossero affidabili e se vi fosse altro da valutare.

Di seguito riportiamo alcuni dei grafici più rilevanti derivati dalle varie modifiche effettuate sui parametri per identificare eventuali pattern interni ai dati.

In figura 4.8 e 4.9 abbiamo le reti prodotte fissando i parametri in modo da mantenere solo i comportamenti più comuni: in 4.8 la rete dei processi lunghi e in 4.9 quella dei processi brevi.
Questi due modelli sono stati prodotti scegliendo dei valori dei parametri che mostrasse solo i pattern più frequenti nei log. Come facilemente visibile entrambe le reti sono praticamente sequenziali nel flusso di esecuzione, l'unica differenza veramente rilevante è determinata dal ciclo in 4.8 tra "CalledPage", "WhoIsInformation" e "PageViewChanged" che mostra come nelle traccia relative a processi di firma lunghi sia il comportamento più comune riaprire il documento più volte, mentre in figura 4.9 relativa a processi svolti tempistivamente la ragione principale della rapidità nella firma risiede nell'esecuzione in una sola apertura del documento dell'intero processo.

Mostriamo invece in figura 4.10 e 4.11 delle reti più generiche che pertanto risultano più complesse e che catturano più comportamenti presenti nel dataset, derivate da un set di parametri meno stringenti.

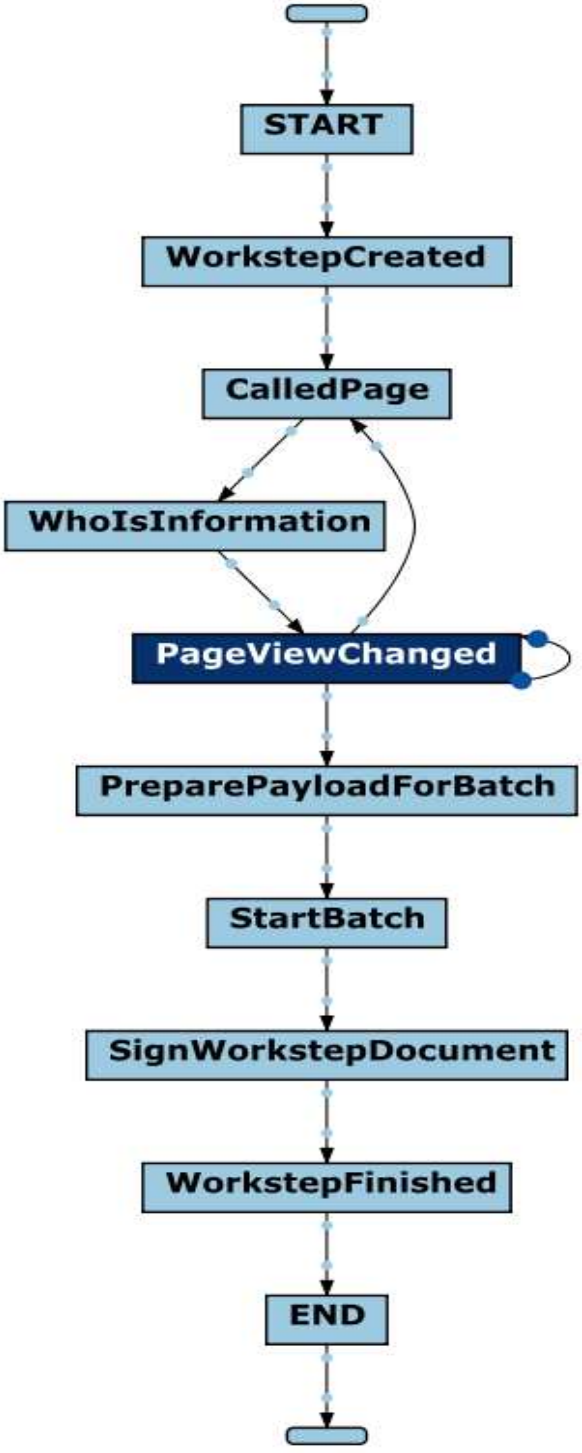

START
WorkstepCreated
CalledPage
WhoIsInformation
PageViewChanged
PreparePayloadForBatch
StartBatch
SignWorkstepDocument
WorkstepFinished
END

Fig. 4.8

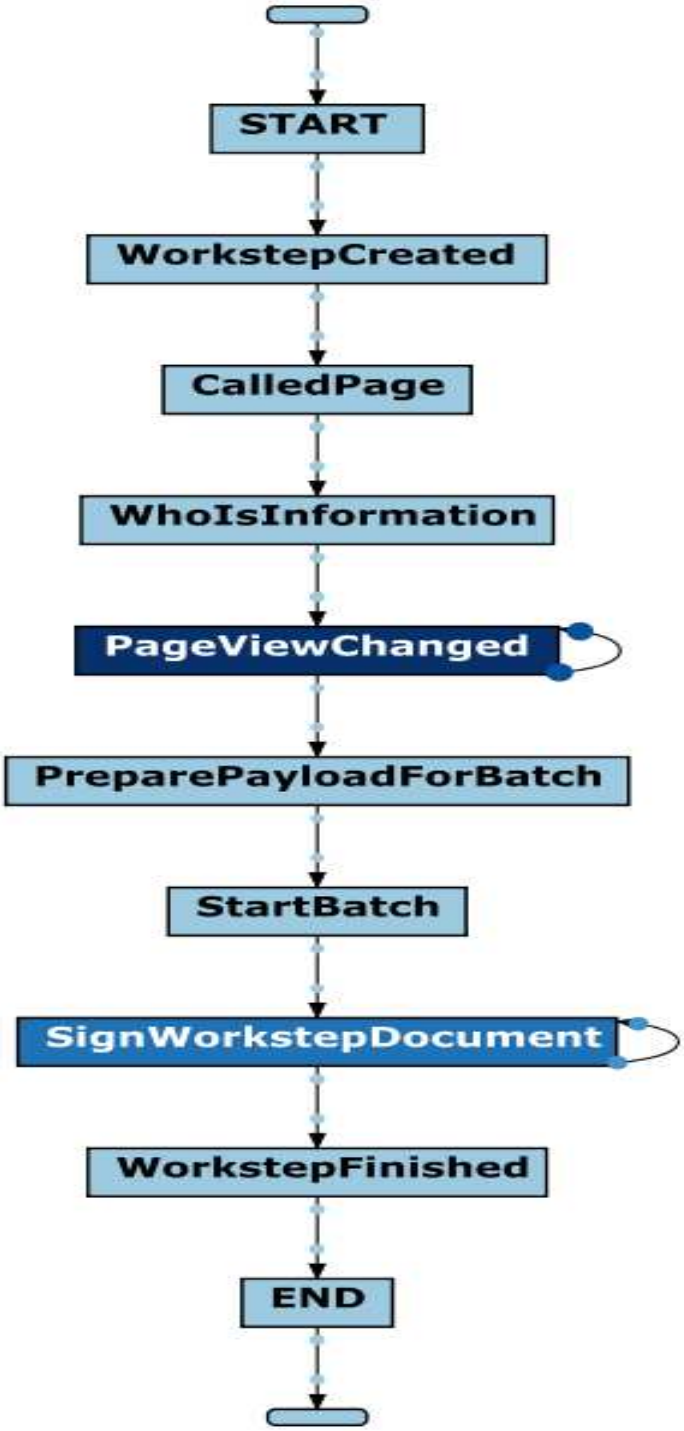

START
WorkstepCreated
CalledPage
WhoIsInformation
PageViewChanged
PreparePayloadForBatch
StartBatch
SignWorkstepDocument
WorkstepFinished
END

Fig. 4.9

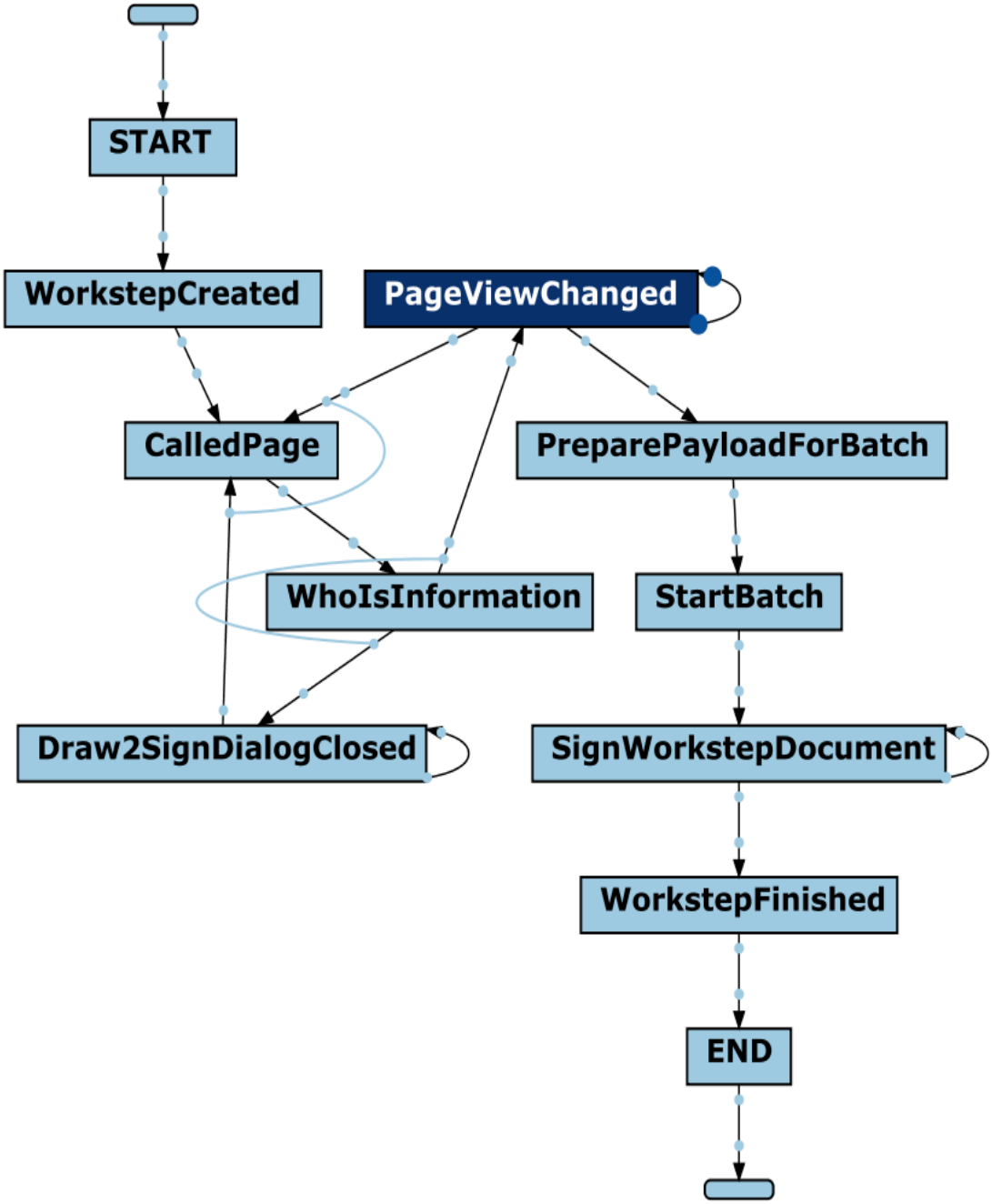

START
WorkstepCreated
PageViewChanged
CalledPage
PreparePayloadForBatch
WhoIsInformation
StartBatch
Draw2SignDialogClosed
SignWorkstepDocument
WorkstepFinished
END

Fig. 4.10

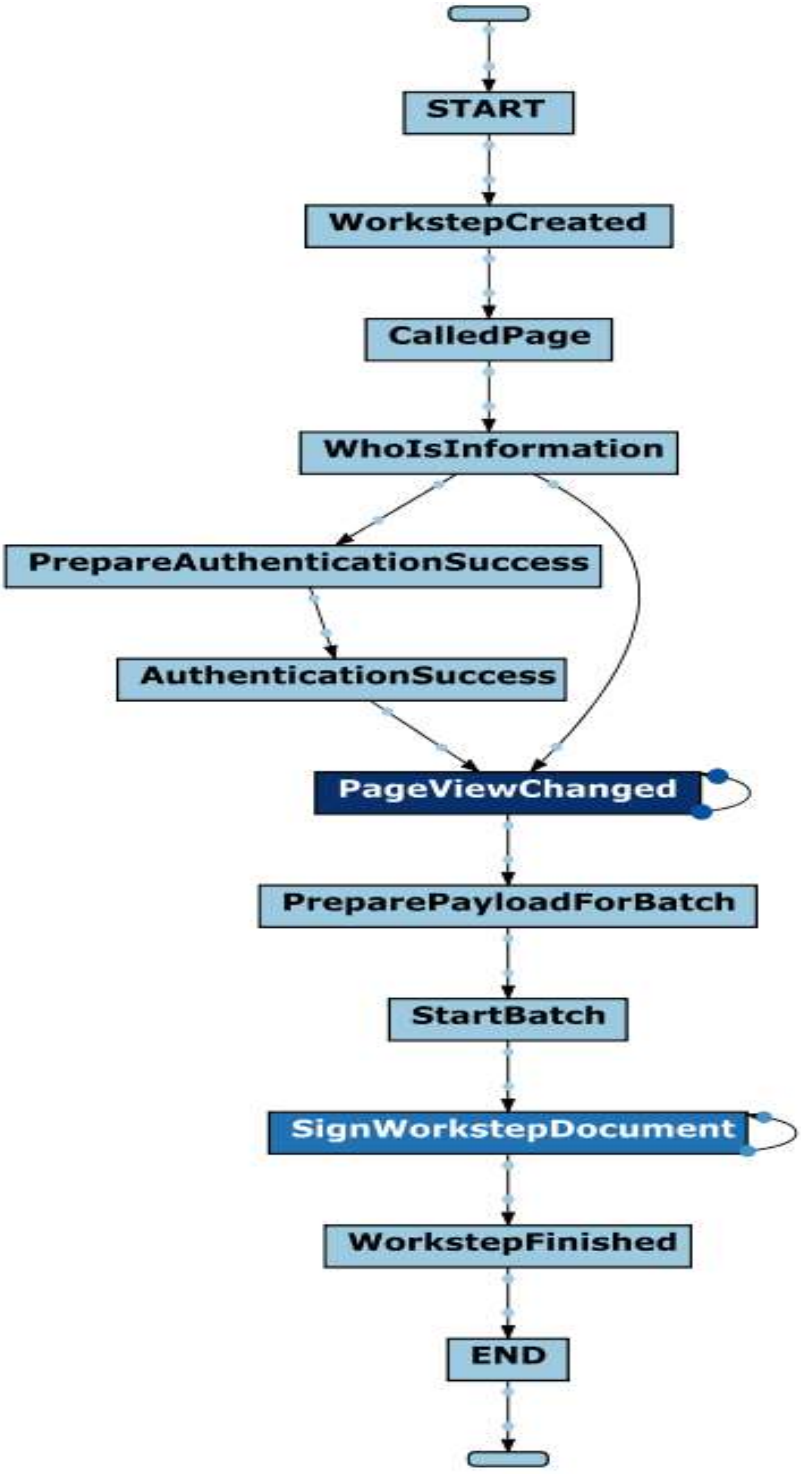

START
WorkstepCreated
CalledPage
WhoIsInformation
PrepareAuthenticationSuccess
AuthenticationSuccess
PageViewChanged
PreparePayloadForBatch
StartBatch
SignWorkstepDocument
WorkstepFinished
END

Fig. 4.11

Risulta facilmente visibile da queste due reti più dettagliate lo stesso fatto presente anche nei diagrammi meno dettagliati: i processi rapidi non hanno riaperture del documento, cosa che hanno invece i processi considerati lunghi.

Altro fatto molto rilevante è la presenza dell'evento "Draw2SignDialogClosed" e di come nel modello risulti che il passaggio alla riapertura del documento passi per questo evento e per l'evento "PageViewChanged".
Inoltre ulteriori variazioni dei parametri non hanno portato alla rilevazione di pattern differenti o particolarmente interessanti internamente al dataset, limitandosi a sporcare le reti con troppi dettagli o a compattarle eccessivamente in un insieme troppo ristretto di eventi.

In ogni caso i risultati evidenziati dall'Heuristic Miner di ProM confermano come siano rilevanti sia la riapertura dei documenti come principale causa del rallentamento nel processo di firma sia la presenza di "Draw2SignDialogClosed" come causa o concausa della riapertura dei documenti.

Al fine di avere maggiore sicurezza nei risultati ottenuti abbiamo avviato la costituzione di altre reti attraverso il fuzzy miner implementato dal tool Disco, di seguito mostriamo in figura 4.12 e 4.13 le reti prodotte.

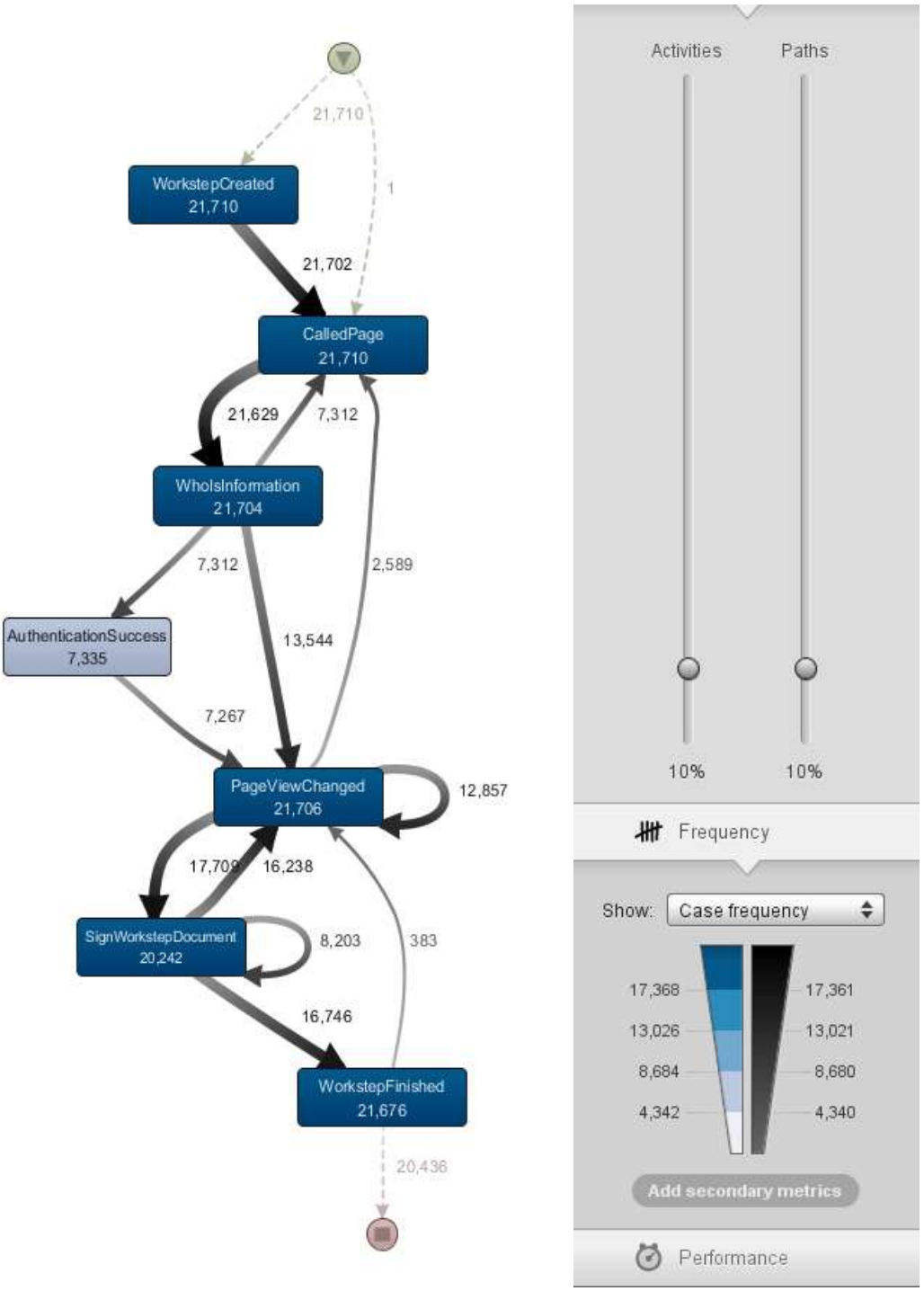

21,710
WorkstepCreated
21,710
1
21,702
CalledPage
21,710
21,629
7,312
WhoIsInformation
21,704
7,312
2,589
AuthenticationSuccess
7,335
13,544
7,267
PageViewChanged
21,706
12,857
17,709
16,238
SignWorkstepDocument
20,242
8,203
383
16,746
WorkstepFinished
21,676
20,436
Activities
Paths
10%
10%
Frequency
Show:
Case frequency
17,368
13,026
8,684
4,342
17,361
13,021
8,680
4,340
Add secondary metrics
Performance

Fig. 4.12

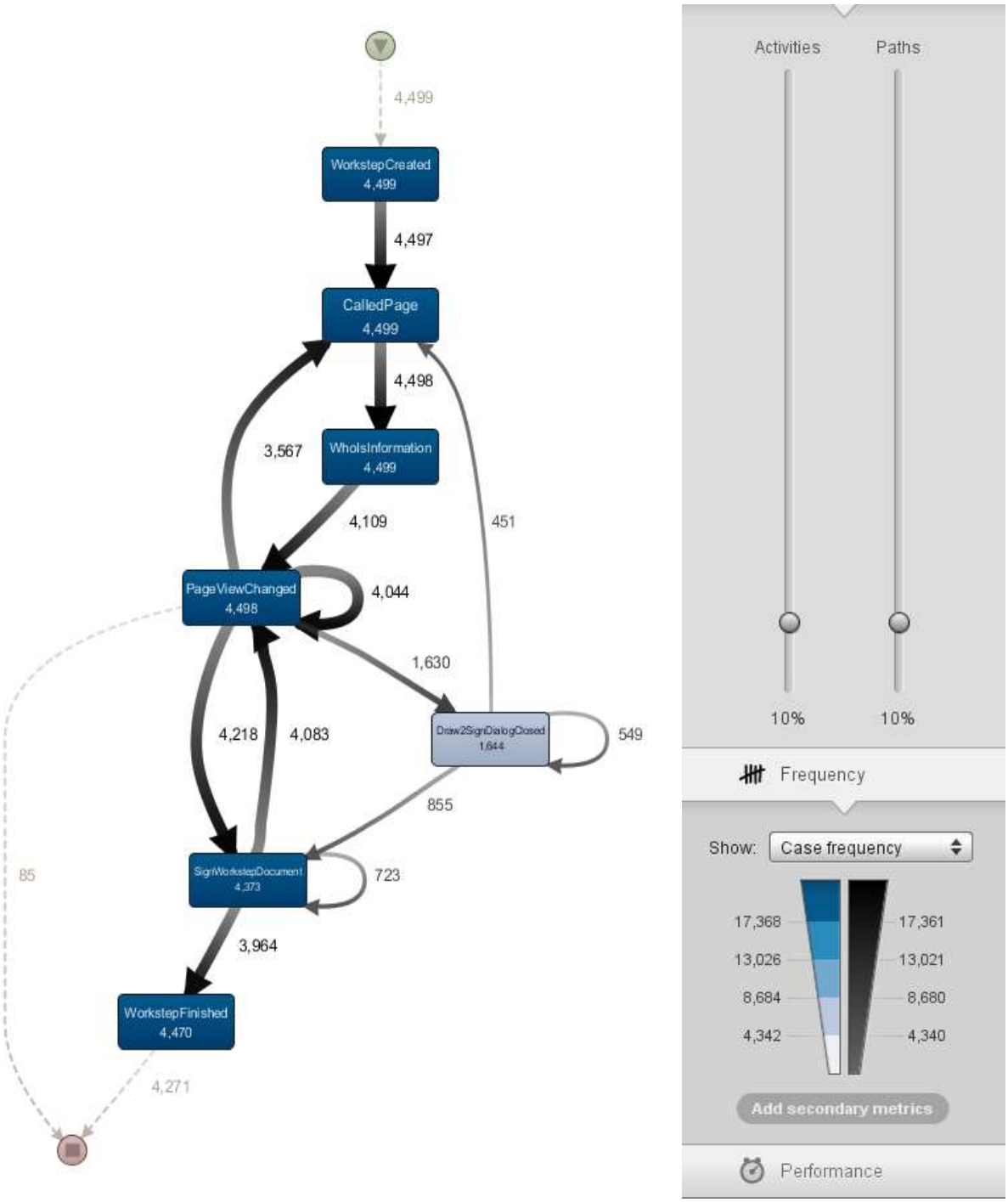

4,499
WorkstepCreated
4,499
4,497
CalledPage
4,499
4,498
WhoIsInformation
4,499
3,567
4,109
451
PageViewChanged
4,498
4,044
1,630
4,218
4,083
Draw2SignDialogClosed
1,644
549
855
85
SignWorkstepDocument
4,373
723
3,964
WorkstepFinished
4,470
4,271
Activities
Paths
10%
10%
Frequency
Show:
Case frequency
17,368
13,026
8,684
4,342
17,361
13,021
8,680
4,340
Add secondary metrics
Performance

Fig. 4.13

La figura 4.12 rappresenta la rete relativa ai processi corti, mentre la 4.13 quella relativa ai processi lunghi. I numeri associati presenti rappresentano in quanti casi (interi log di una singola firma) è avvenuta quella transizione, per gli archi, o quell'evento, per i riquadri.
La valorizzazione dei parametri dell'algoritmo è stata determinata avendo come obiettivo la massima semplicità nella lettura dei modelli pur volendo mantenere tutte le informazioni inerenti all'impatto che l'evento "Draw2SignDialogClosed" ha nel flusso di esecuzione delle firme. È possibile notare come nei processi corti il suddetto evento non risulta considerato dal modello, mentre tra i devianti ricopre un ruolo drasticamente più rilevante venendo sia mantenuto come evento, sia avendo un ramo che ritorna a "CalledPage" ovvero che conduce alla riapertura del documento.

Si sono altresì valutati i grafici che manteneva tutti i possibili flussi di lavoro sia dei processi corti che dei processi lunghi, non li riportiamo poiché estremamente caotici e illeggibili senza l'ausilio del tool Disco. Nella seguente tabella però riportiamo delle misure inerenti al nodo "Draw2SignDialogClosed" nelle due reti distinte dei processi lunghi e brevi.

	Draw/Totale	ToCall/Draw	ToSign/Draw	ToViewChanged/Draw	ToDraw/Draw
BREVI	0.111	0.095	0.68	0.375	0.264
LUNGHI	0.365	0.274	0.663	0.513	0.333

La tabella mostra il risultato di vari rapporti tra il numero di casi in cui sono avvenute determinate transizioni. Naturalmente la somma dei risultati su una riga non deve essere uguale a 1 dal momento che ciascuna transizione non è in alcun modo esclusiva con le altre su un singolo caso. Riportiamo ora il significato dei rapporti nelle intestazioni:

- Draw/Totale:

  numero di casi in cui è presente l'evento "Draw2SignDialogClosed" e il numero totale dei casi presenti nella rete considerata

- ToCall/Draw:

  numero di casi in cui si attiva l'arco dall'evento "Draw2SignDialogClosed" all'evento "CalledPage" e il numero di casi in cui è presente l'evento "Draw2SignDialogClosed"

- ToSign/Draw:

  numero di casi in cui si attiva l'arco dall'evento "Draw2SignDialogClosed" all'evento "SignWorkstepDocument" e il numero di casi in cui è presente l'evento "Draw2SignDialogClosed"

- ToViewChanged/Draw:

  numero di casi in cui si attiva l'arco dall'evento "Draw2SignDialogClosed" all'evento "PageViewVhanged" e il numero di casi in cui è presente l'evento "Draw2SignDialogClosed"

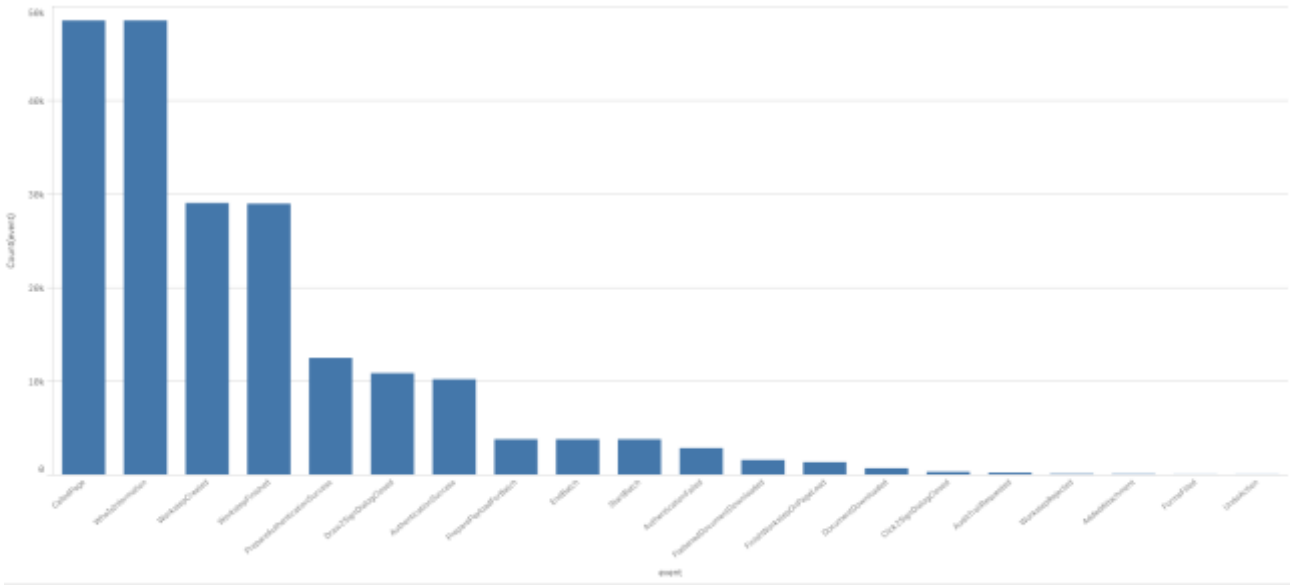
- ToDraw/Draw:

  numero di casi in cui si attiva l'arco dall'evento "Draw2SignDialogClosed" all'evento stesso e il numero di casi in cui è presente l'evento "Draw2SignDialogClosed"

Dalla tabella risulta subito come quasi tutte le metriche selezionate siano più alte nei processi lunghi, a eccezione di "ToSign/Draw" la quale è comparabile tra i due casi e rappresenta quanto le firme vengano effettivamente apposte dopo aver chiuso almeno una volta l'interfaccia di firma, ovvero una forma di probabilità di rapido successo del task dal momento che si chiudi almeno una volta il pop-up di firma. Per le altre metriche otteniamo che sembra esserci una correlazione positiva tra il fatto che il processo possa essere categorizzato come lungo e la presenza dell'evento "Draw2SignDialogClosed" o a che da lì la transizione sia verso attività altre rispetto all'effettiva firma del campo.

Infine per poter essere certi dell'impatto dell'evento "Draw2SignDialogClosed" sull'effettiva chiusura e riapertura del documento si è ritenuto opportuno realizzare uno script python che derivasse la percentuale su tutto il dataset senza distinguere tra processi lunghi e brevi in cui l'evento conducesse a una riapertura del documento: in 4631 pratiche vi è la presenza dell'evento e in 1697 pratiche vi è una chiusura e riapertura del documento a seguito dell'evento "Draw2SignDialogClosed" pertanto vi è una confidenza del 36,6% nel dire che dal momento che un utente chiude una

finestra di dialogo esso poi chiuderà anche il documento per firmarlo in seguito. Inoltre si è anche scoperto che all'interno del dataset i dati relativi alla firme attraverso la procedura che prevede il disegno della firma sono tutti relativi a una sola compagnia tra le considerate e che in tutto il dataset l'analogo all'evento "Draw2SignDialogClosed" per la firma solo con click, ovvero "Click2SignDialogClosed" è presente drasticamente meno, figura 4.14.



4.14

L'ultima questione che ci siamo curati di valutare è stata, attraverso delle analisi su dashboard interattive, se vi fossero delle azioni che fungessero da collo di bottiglia sulla solerzia del processo al di là delle aperture e chiusure.

Per tale scopo è risultato sufficiente, dopo il necessario preprocessing, costituire il grafico a barre con i dati pre processati per rimuovere le pause delle chiusure del documento così da avere solo durate relative all'effettiva durata delle azioni in figura 4.15.

Nella parte superiore si mostra la numerosità degli eventi divisi per tipo così da poter dare maggior garanzie in termini statistici per la validità della durata media di quel tipo di eventi.

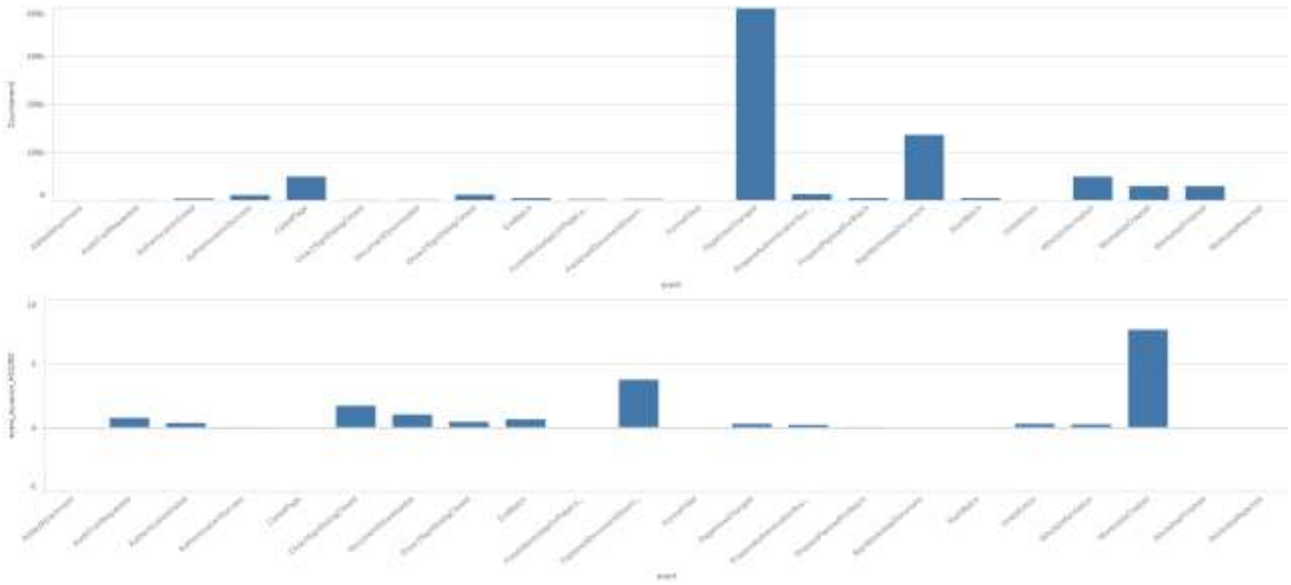


Fig. 4.15

Si nota facilmente come non vi siano effettivamente degli eventi che avendo un'alta numerosità siano anche altamente impattanti dal punto di vista dei tempi: a eccezione

di “WorkstepCreated” che però non è pertinente come sappiamo al puro processo di firma dell’utente.

Inoltre nel grafico in figura 4.16 si riporta il confronto tra le durate effettiva dei processi di firma, nella grafico nella parte superiore della figura, e le durate dei processi di firma nel caso in cui venissero compattati eliminando le pause in cui si chiude il documento, nella parte inferiore.

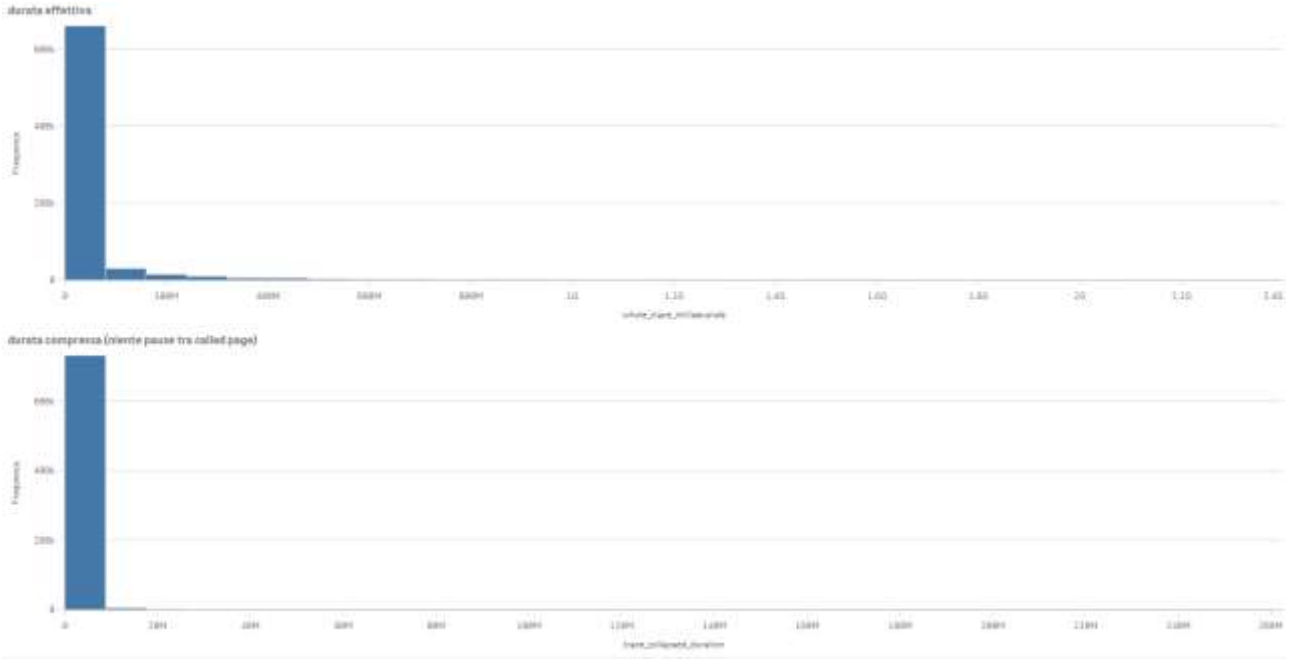


Fig. 4.16

Si nota agevolmente come l’effettiva durata sia tutta frutto del ritardo tra una chiusura e la successiva riapertura, essendo che i nel grafico inferiore in 4.16, quindi

quello relativo alle durate collassate, le durate sono veramente esegue se confrontate con il grafico delle durate effettive.

5. Conclusione e lavori futuri

L'insieme alle analisi svolte induce a pensare come effettivamente il problema principale che ostacola la sottoscrizione tempestiva dei documenti sia la chiusura e riapertura da parte degli utenti del documento per poi firmarlo in un momento successivo. Un fatto altresì importante però risulta essere l'impatto che la necessità di disegnare la propria firma ha sulla probabilità che il documento venga chiuso e riaperto in un secondo momento: viene inevitabilmente da pensare che il compito di disegnare la propria firma attraverso un'interfaccia digitale risulti ostico e frustrante per la maggior parte degli utenti. Inoltre dal momento che la forza legale della firma non è strettamente collegata a questo aspetto di disegno della stessa potrebbe essere saggio, volendo produrre una raccomandazione per il servizio, evitare di basarsi su firme che richiedano un disegno.

Gli strumenti utilizzati si sono rilevati più che appropriati al dominio applicativo in questione e pertanto delle successive direzioni di ricerca potrebbero essere relative all'arricchimento degli eventi con i dati relativi ai documenti, specializzando di fatto alcune categorie di azioni in relazione a ciò cui fanno riferimento. Questo arricchimento potrebbe richiedere la messa in atto di varie tecniche che lavori in modo sinergico e ausiliario al process mining, quali il natural language processing e il clustering.

Bibliografia

[1] Chapman, Pete & Clinton, Julian & Kerber, Randy & Khabaza, Thomas & Reinartz, Thomas & Shearer, Colin & Wirth, Rüdiger. (1999). CRISP-DM 1.0 step-by-step data mining guide.

[2] Song, Minseok & Günther, Christian & Aalst, Wil M. P.. (2008). Trace Clustering in Process Mining. Lecture Notes in Business Information Processing. 17. 109-120. 10.1007/978-3-642-00328-8_11.

[3] Mannhardt, Felix & de Leoni, Massimiliano & Reijers, Hajo. (2017). Heuristic Mining Revamped: An Interactive, Data-aware, and Conformance-aware Miner.

[4] Günther, Christian & Aalst, Wil M. P.. (2007). Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics. 4714. 328-343. 10.1007/978-3-540-75183-0_24.