Abstract

Il termine process mining racchiude un insieme di tecniche e metodolgie utilizzate nell'ambito del Business Process Management per l'analisi e caratterizzazione di processi a supporto di sistemi transazionali (e.g., Workflow Management (WF), Sistemi di Supply Chain Management (SCM), Applicazioni di Business to Business (B2B), Sistemi di Customer RelationShip Management) [114].
L'idea alla base del process mining consiste nel costruire modelli significativi del comportamento di processi a partire dall'analisi dei loro log di esecuzione. Tali modelli sono particolarmente utili in ambito industriale in quanto consentono di monitorare, ottimizzare o completamente ridefinire complessi schemi di processo a partire dall'analisi del loro reale funzionamento.
Nel lavoro di tesi presentiamo una serie di algoritmi, metodi e sistemi espressamente progettati per supportare le odierne tecniche di process mining nell'affrontare alcuni aspetti peculiari dell'analisi di log in contesti reali. In dettaglio:

- Definiamo una nuova tecnica di *outlier detection* per il process mining a supporto delle tradizionali tecniche euristiche [81,82,123] definite in letteratura. Le tecniche euristiche si propongono di costruire modelli ragionevoli a partire da log affetti da rumore ed incompletezza. Tali tecniche utilizzano dei threshold forniti dall'utente per definire la frequenza di esecuzione minima al di sotto della quale una sequenza di attività del processo memorizzata nel log è considerata rumorosa. Tale strategia, tuttavia, potrebbe essere in alcuni casi fuorviante a causa dell'alto grado di parallelismo che tipicamente si verifica in processi reali durante l'esecuzione delle attività. In tale scenario, infatti, sequenze di esecuzione non frequenti potrebbero comunque derivare da modalità di funzionamento ricorrenti del processo. Sulla base di questa intuizione, la tecnica proposta identifica le anomalie e gli errori potenzialmente presenti in un log non solo come deviazioni dalle sequenze di attività registrate ma anche come deviazioni dal comportamento prescritto dai *frequent behavioral pattern* di un certo modello del processo, eventualmente sconosciuto. Tale peculiarità garantisce una buona robustezza della tecnica al variare della concentrazione di outlier nei log.

- Definiamo un insieme di tecniche efficienti atte a velocizzare gli attuali metodi di process mining multi-relazionale [43, 73]. A differenza dei tradizionali approcci basati sull'analisi delle sole sequenze di attività memorizzate nei log, gli approcci multi-relazionali esplorano l'intrinseca struttura relazionale del log alla ricerca di interessanti relazioni logiche tra le sequenze di attività di un processo e i dati "non strutturali" associati all'environment in cui il processo stesso è eseguito. L'analisi di tali relazioni consente di ottenere modelli di ottima qualità in grado di catturare comportamenti complessi non deducibili dalla sola analisi delle sequenze di attività [37]. Nonostante le loro potenzialità, le tecniche multi-relazionali per il process mining mostrano considerevoli limiti di scalabilità dovuti all'utilizzo di lenti risolutori ILP per l'inferenza dei modelli [105, 132]. Le tecniche proposte in questa tesi derivano dalle nostre esperienze di ricerca nella Teoria delle Basi di Dati [41] e si basano su sofisticati metodi di query answering in grado di risolvere efficientemente query complesse usando sia meccanismi di basso livello implementati nei DBMS che le proprietà strutturali delle interrogazioni. In particolare, proponiamo un'estensione della tecnica *hypertree decomposition* [45] espressamente definita per il setting multi-relazionale e definiamo delle efficienti implementazioni basate su DBMS della tecnica `TupleIDPropagation` [132] che rappresenta attualmente la strategia di riferimento nel panorama multi-relazionale. Tale tecnica, infatti, si è dimostrata superiore in termini di scalabilità ai comuni sistemi ILP utilizzati per il mining multi-relazionale ed è adatta all'analisi di dataset di grandi dimensioni [132].

- Definiamo l'architettura di PROMETHEUS, un sistema completo per applicazioni di process mining. Il sistema nasce come naturale evoluzione di ProM [116], il più conosciuto ed utilizzato framework per process mining, del quale riprende e migliora alcuni aspetti salienti aggiungendo al contempo innovative caratteristiche espressamente progettate per l'analisi di log in contesti reali. PROMETHEUS introduce l'innovativo concetto di *flow of mining* che consente di gestire in maniera flessibile ed intuitiva più task di analisi contemporaneamente. Secondariamente, PROMETHEUS consente la definizione di plug-in interattivi, completamente configurabili e customizzabili con eventuali componenti grafici di supporto all'analisi. Infine, PROMETHEUS introduce un approccio stream based per l'acquisizione e l'analisi di log di grande dimensione. Tale caratteristica differenzia

sostanzialmente PROMETHEUS da tutti gli altri sistemi per process mining (che propongono invece una gestione in memoria dei dati) e rende il sistema affidabile in complessi scenari industriali dove è necessario eseguire algoritmi di mining su terabyte di dati.