UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DIPARTIMENTO DI INFORMATICA

Dottorato di ricerca in Informatica e Matematica
XXXI ciclo

Progetto di ricerca

Dottorando: Dott. Michele Spagnoletta
Tutor: Prof. Michelangelo Ceci
Tutor aziendale: Dott. Felice Vitulano

Coordinatore
Prof. Donato Malerba

Firma del dottorando _____

Firma del tutor _____

Firma del tutor aziendale _____

1) Titolo della ricerca:

Analisi dei processi in contesti aziendali con l'utilizzo di tecnologie di Big Data Analytics

2) Area nella quale si inquadra la ricerca:

Data Mining e Machine Learning, Process Mining, Big Data

3) Obiettivi della ricerca:

Modellare, monitorare e migliorare i processi estraendo conoscenza dai log (informazioni legate all'esecuzione dei processi nel mondo reale) e cogliere la strategia vincente per migliorare la qualità dei processi e ridurre i costi. Un log, infatti, può contenere eventi di vario tipo: un utente che ritira del denaro contante a uno sportello automatico, un operatore di call center che prende in carico una chiamata e la conclude risolvendo o meno un problema, un utente che naviga nel web sono tutti scenari in cui l'azione è tracciata. L'obiettivo è cercare di sfruttare questi dati in modo significativo: per esempio per fornire suggerimenti, fare predizione, identificare colli di bottiglia, prevedere problemi, registrare violazioni di regole, raccomandare contromisure e dare "forma" ai processi.

A tal fine, sono tipicamente adottate tecniche di data mining che, data la loro specificità, prendono il nome di tecniche di process mining. Tali tecniche, sebbene siano state proficuamente utilizzate in passato in ambito aziendale, riescono solo limitatamente a gestire e trarre profitto dall'enorme quantità di dati che i processi aziendali sono in grado di generare in modo (semi-)automatico. Il dottorando, quindi, durante il Dottorato di Ricerca, intende concentrare l'attenzione, proporre e studiare soluzioni di analisi dei dati che siano in grado di affrontare e risolvere i problemi e le sfide classiche del process mining utilizzando tecniche e tecnologie di Big Data Analytics.

Tali tecniche si inquadreranno appieno nell'ottica "Big Data" che rappresenta una opportunità unica per le organizzazioni di capitalizzare dati raccolti in maniera massiva e provenienti da molteplici fonti eterogenee che includono i dati strutturati tradizionali così come nuovi fonti non strutturate come i registri, i dati di strumentazione, dati di rete, videosorveglianza RSS, informazioni geospaziali, dati sociali e molto altro.

Considerando ad esempio l'ambito della sicurezza aziendale, l'utilizzo di tecnologie di Big Data Analytics consentono di migliorare la sorveglianza sfruttando diverse fonti e tipi di dati, come ad esempio internet, satellite, video e audio ma anche informazioni proveniente da social network, informazioni geospaziali, testi, immagini, video e dati vocali sono necessari per identificare e prevenire la criminalità.

Il dottorando si concentrerà principalmente sulla sintesi di nuove soluzioni di data mining parallelo e distribuito per compiti classici di process mining che offrono la possibilità di elaborare enormi quantità di log e, in generale, di dati eterogenei tra loro, che sono impossibili da elaborare con tecniche classiche di data mining.

4) Motivazioni della ricerca

Le motivazioni della ricerca nascono dalla crescente disponibilità dei dati che forniscono informazioni dettagliate sulle esecuzioni dei processi e dalla necessità di migliorare e supportare i processi di business in contesti sempre più competitivi e in rapida evoluzione (come quello di un call center).

Nonostante la presenza di approcci già noti in letteratura per lo specifico obiettivo di ricerca, bisogna considerare alcune sfide che devono essere affrontate:

Scoperta, Fusione e Pulizia di dati di eventi

I log reali possono essere influenzati dal rumore in quanto possono contenere informazioni incomplete o errate, secondo il settore specifico. Per esempio, in caso di dati raccolti dai sensori, è possibile avere problemi di trasmissione di dati e malfunzionamento di sensori che impediscono i dati di raggiungere il luogo immagazzinamento. L'event log può contenere outliers ed eventi a differente livello di granularità.
La big data analytics migliora notevolmente non solo la scoperta di eventuali outlier ma consente anche l'interpolazione nel sostituire dati mancanti e outlier.

Manipolazione log complessi e con caratteristiche diverse

I sistemi raccolgono automaticamente una massiccia quantità di dati che vengono memorizzati in sempre più crescenti file / database. Con l'affermarsi dei big data si avranno log sempre più complessi e con caratteristiche eterogenee vista la provenienza da fonti diverse, ma questo rappresenta anche un vantaggio in quanto possono fornire informazioni più dettagliate e provenienti da diverse fonti (e, quindi, con diversi punti di vista) sulla storia dei processi. Le tecnologie di big data analytics, inoltre, consentono di memorizzare, trattare ed elaborare tutti i dati in modo efficiente.

Trattamento del concept drift

Il termine concept drift indica il fatto che il processo che si sta analizzando è in continua evoluzione. Il cambiamento di un processo può essere legato a cambiamenti periodici/ stagionali oppure perché cambiano le condizioni al contorno. In ogni caso i cambiamenti impattano sui processi. La big data analytics in tal caso è di vitale importanza in quanto riesce a individuarli ed analizzarli.

Distribuzione dei dati

Per motivi di privacy o per ridurre i costi di trasmissione, i dati possono essere, soprattutto nelle grandi organizzazioni, distribuiti. In questi casi, lo spostamento di dati per un sito centrale per lavori di estrazione può diventare un problema insormontabile. I dati possono essere distribuiti su diverse sorgenti informative; essi possono essere identificati in modo diverso.

Mediante l'utilizzo delle tecnologie delle basi di dati NoSQL distribuite (e.g., HBASE) e del big data analytics (e.g. Spark) è possibile gestire, manipolare, distribuire e analizzare set di dati di dimensioni tali che in database tradizionali non possono essere utilizzati.

Supporto alle decisioni

Il Process Mining non deve essere più ristretto ad un'analisi off-line ma può essere esteso anche per attività di supporto decisionale in tempo reale. Il supporto alle decisioni prevede tre tipi di attività: rilevazione, predizione, raccomandazione. In tal caso, tecniche di big data analytics, lavorando su una maggiore quantità di dati e su dati eterogenei, sono in grado di supportare analisi con *confidenza* statistica maggiore.

I sistemi esistenti si pongono e affrontano solo alcune di queste sfide. L'obiettivo del progetto di ricerca è quello di affrontare in maniera sistematica e combinata le problematiche elencate utilizzando tecnologie di big data analytics.

5) Stato dell'arte

Lo scopo del Process Mining è quello di dare "forma" ai processi [1], mentre il punto di partenza per qualsiasi tecnica di Process Mining è un log degli eventi.

Gli event log possono essere usati per eseguire tre tipi di Process Mining:

1. DISCOVERY: una tecnica di discovery prende in input un event log e produce un modello senza utilizzare alcuna informazione a priori. Il process discovery è la più importante tecnica di Process Mining, e molte organizzazioni che ne hanno avuto esperienza trovano stupefacente come queste tecniche possano effettivamente descrivere processi reali solamente basandosi su esempi di esecuzione. Le tecniche di discovery [2] prendono in input un event log e producono un modello che è tipicamente un modello di processo (per esempio una rete di Petri, un diagramma UML delle attività). Tuttavia, il modello può anche descrivere altre prospettive (come per esempio una social network).
2. CONFORMANCE CHECKING: un modello di processo preesistente è confrontato con informazioni (relative allo stesso processo) estratte da un event log. Il conformance checking può essere usato per verificare se ciò che accade nella realtà (come risulta dai log) è conforme al modello e viceversa. Le tecniche prendono in input un event log e un modello. L'output consiste in una serie di informazioni diagnostiche che mostrano le differenze tra il modello e il log.
3. ENHANCEMENT: l'idea è quella di estendere o migliorare un modello di processo esistente usando informazioni circa il processo contenute nei log. Le tecniche richiedono un event log e un modello in input. L'output è il modello stesso migliorato o esteso.

In questi tipi di process mining si presume che processo di estrazione avviene **offline**. I processi sono analizzati *a posteriori* per valutare come possano essere migliorati o estesi. Al contrario, le tecniche di *Operational Support* sono utilizzate in impostazioni **online**. Dato un modello di processo costruito su alcuni log di eventi e una traccia parziale, le tecniche di Operational Support possono essere utilizzate per rilevare la deviazione in fase di esecuzione (Detect), prevedere il tempo di elaborazione residuo (Predict) e per raccomandare l'attività successiva (Recommend). Tipicamente, gli algoritmi classici presentati in letteratura per l'Operational Support, (1) costruiscono un modello di processo in forma di sistemi di transizione [3] o Petri-Nets [4,5], (2) ri-analizzano i log per estendere il modello con informazioni temporali e statistiche aggregate [6], o, infine, (3) apprendono un modello di regressione o un modello di classificazione per supportare le attività di previsione e di raccomandazione. Tuttavia, come osservato in [7], questi metodi di Operational Support si adattano naturalmente a casi in cui i processi sono molto ben strutturati (cioè sono in perfetta armonia con un qualche schema predefinito), per i log reali, invece, che soffrono di problemi legati a:

- "incompiutezza" (vale a dire il modello rappresenta solo una piccola frazione del possibile comportamento a causa del gran numero di alternative),
- "rumore" (cioè, i registri contenenti attività eccezionali / infrequente che non dovrebbero essere incorporate nel modello),
- "overfitting" e "under-fitting"

e portano a modello di spaghetti-like, che sono piuttosto inutili nella pratica.

Altri approcci, come sistemi di intelligenza computazionale [8], che superano questi problemi, tendono ad essere inefficienti e, pertanto, hanno problemi nello scalare in caso di una grande quantità di attività che sono correlati tra loro (mediante precedenza / dipendenze causalità).

In questo progetto si intendono superare tali problemi mediante tecniche di big data analytics che sono anche in grado di scalare bene su dati di grandi dimensione e distribuiti. Particolare enfasi sarà rivolta alle applicazioni.

In letteratura, sono state proposte molte tecniche di Operational Support che lavorano in contesti applicativi diversificati. Per esempio, l'articolo [9] riporta l'applicazione di tecniche di process mining nel reparto di oncologia di uno dei principali ospedali olandesi: il log utilizzato contiene

eventi che si riferiscono a 376 diverse attività. Un altro interessante caso [10] riguarda l'applicazione di tecniche di process mining combinate con tecniche di classificazione nel contesto dei processi aziendali: l'obiettivo era quello di scoprire diversi scenari di gestione (ognuno descritto con un modello dettagliato di processo) e, inoltre, evidenziare le correlazioni tra tali casi d'uso e alcune proprietà non strutturali dei processi stessi (es. provenienza/destinazione, dimensione). Tecniche di process mining sono anche state utilizzate per generare un modello di processo per i lavori pubblici in Olanda [11].

Tuttavia, c'è da sottolineare che tutte queste tecnologie non sono state mai investigate mediante l'utilizzo di tecnologie di Big Data Analytics. L'unico lavoro che prevede algoritmi di data mining distribuito per l'analisi di log è stato presentato in [12], dove tuttavia si propongono solo soluzioni atte a descrivere i dati e non di Operational Support.

6) Approccio al problema

L'approccio al problema prevede (le fasi non sono ordinate temporalmente):

- Una pianificazione ed una giustificazione dell'attività di pianificazione stessa
- L'interrogazione dei sistemi informativi, esperti di dominio e manager per ricavare dati, modelli, obiettivi e domande a cui è necessario successivamente rispondere. La comprensione dei dati che si hanno a disposizione e del dominio di interesse rislta fondamentale in questo caso.
- Individuazione delle tecniche migliori di process mining per i problemi individuati. Sintesi di nuovi algoritmi per rispondere meglio alle esigenze emerse.
- Filtraggio dei log e adattamento sulla base del modello (eliminando attività rare o istanze anomale, inserendo eventi mancanti). Correlazione degli eventi appartenenti alla stessa istanza di processo. Individuazione di eventuali fonti esterne di dati. in modo da estendere il modello per il flusso di controllo con altre prospettive (ad esempio, tempi, risorse normative).
- Progettazione e implementazione degli algoritmi proposti. Le soluzioni proposteconsentiranno di gestire, manipolare e analizzare set di dati distribuiti, eterogenei e di grandi dimensioni.

- Valutazione sperimentale degli approcci proposti, in modo da individuare criticità e raffinare i metodi proposti.
- Pubblicazione dei risultati scientifici ottenuti.
- Messa in opera delle soluzioni proposte in contesti e applicazioni reali.

7) Ricadute applicative

Negli ultimi dieci anni, le tecniche di process mining sono state utilizzate in più di 100 organizzazioni inclusi comuni (per es., Alkmaar, Zwolle, Heusden in Olanda), agenzie governative (per es., Rijkswaterstaat, Centraal Justitieel Incasso Bureau), banche (per es., ING Bank), ospedali (per es., Catharina hospital ad Eindhoven), multinazionali (peres., Deloitte), industrie manifatturiere e loro clienti (per es., Philips, ASML, Ricoh, Thales). Questa diffusione dimostra l'ampio numero di contesti nei quali è possibile applicare il process mining.

Nel lavoro previsto nell'ambito del progetto formativo si intendono sintetizzare nuove soluzioni di process mining per i seguenti contesti applicativi (lista non vincolante e non limitativa):

- Call center: Nell'ultimo periodo si ha una crescita di call center sia per assistenza che per vendita di servizi e prodotti. L'utilizzo del process mining consentirebbe di creare un motore di operation intelligence che mappa tutto ciò che succede al fine di ottimizzare il lavoro, far scattare degli alert, fare un'analisi predittiva come il suggerimento della prossima attività da svolgere dall'operatore, la stima del tempo rimanente per il completamento di un processo, le criticità previste, il conformal prediction, i tempi di presi in carico e risposta.
- Telefonia: Sempre più aziende di telefonia hanno la necessità di sfruttare maggiormente le informazioni che esse stesse producono come ad esempio i dati di navigazione dei propri utenti in forma anonima e i dati relativi alle chiamate effettuate. Si potrebbero pertanto applicare tecniche di process mining ai log di navigazione per anticipare ad esempio le richieste da parte degli utenti e, di conseguenza, ottimizzare le richieste delle risorse, o comprendere l'eventuale motivo di abbandono di un utente verso un altro gestore in modo da anticipare eventualmente altre perdite di clienti proponendo servizi ad hoc.

8) Riferimenti bibliografici

1. AA.VV. Manifesto del Process Mining. IEEE Task Force on Process Mining. http://www.win.tue.nl/ieeetfpm/lib/exe/fetch.php?media=shared:pmm-italian-v2.pdf

2. W.M.P. van der Aalst. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, Berlin, 2011. Sito uffciale del libro: http://www.processmining.org/book/

3 . van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes, 1st edn. Springer Publishing Company, Incorporated (2011)

4. Dongen, B., Busi, N., Pinna, G., Aalst, W.: An Iterative Algorithm for Applying the Theory of Regions in Process Mining. In: Proceedings of the Workshop on Formal Approaches to Business Processes and Web Services, pp. 36–55 (2007)

5. Carmona, J., Cortadella, J., Kishinevsky, M.: A Region-Based Algorithm for Discovering Petri Nets from Event Logs. In: Dumas, M., Reichert, M., Shan, M.-C. (eds.) BPM 2008. LNCS, vol. 5240, pp. 358–373. Springer, Heidelberg (2008)

6. van der Aalst, W.M.P., Schonenberg, M.H., Song, M.: Time prediction based on process mining. Inf. Syst. 36(2), 450–475 (2011)

7. Folino, F., Greco, G., Guzzo, A., Pontieri, L.: Mining usage scenarios in business processes: Outlier-aware discovery and run-time prediction. Data Knowl. Eng. 70(12), 1005–1029 (2011)

8. Medeiros, A.K., Weijters, A.J., Aalst, W.M.: Genetic process mining: An experimental evaluation. Data Min. Knowl. Discov. 14(2), 245–304 (2007)

9. R. S. Mans, M. H. Schonenberg, M. Song, W. M. P. van der Aalst and P. J. M. Bakker, Application of Process Mining in Healthcare A Case Study in a Dutch Hospital Communications in Computer and Information Science, 1, Volume 25, Biomedical Engineering Systems and Technologies, Part 4, Pages 425-438 Springer-Verlag, Berlin, 2009.

10. F. Folino, G. Greco, A. Guzzo, L. Pontieri. Mining usage scenarios in business processes: Outlier-aware discovery and run-time prediction. Data Knowledge Engineering, Volume 70, nr 12, Pages 1005-1029, 2011. Elsevier

11. W.M.P. van der Aalst, H.A. Reijersa, A.J.M.M. Weijters, B.F. van Dongen, A.K. Alves de Medeiros, M. Song, H.M.W. Verbeek Business process mining: An industrial application Journal Information Systems, Volume 32 Issue 5, July, 2007, Pages 713-732, Elsevier Science Ltd. Oxford, UK

12. APPICE A, CECI M, TURI A, MALERBA D (2011). A parallel, distributed algorithm for relational frequent pattern discovery from very large data sets. INTELLIGENT DATA ANALYSIS, vol. 15, p. 69-88, ISSN: 1088-467X, doi: 10.3233/IDA-2010-0456

9) Fasi del progetto

Anno 1°: studio della letteratura, dello stato dell'arte e del materiale di ricerca di base:

- **Attività 1A:** studio approfondito di aspetti teorico-formali;
- **Attività 1B:** approfondimento delle tematiche relative alle tecniche di process mining di grandi mole di dati
- **Attività 1C:** ricerca, studio, analisi di applicabilità in contesti reali che emergono in ambito aziendale;
- **Attività 1D:** partecipazione a scuole internazionali inerenti all'attività e agli obiettivi previsti.

Anno 2°: sintesi, realizzazione e implementazione di metodi:

- **Attività 2A:** confronto con l'attività svolta da gruppi di ricerca con obiettivi affini;
- **Attività 2B:** sintesi, progettazione e implementazione di metodi che soddisfino gli obiettivi previsti;
- **Attività 2C:** valutazione dei metodi realizzati, confronto con approcci esistenti e pubblicazione dei risultati conseguiti in riviste e conferenze internazionali.

Anno 3°: applicazione ai domini applicativi individuati e stesura della tesi di dottorato:

Attività 3A: stage presso la sede dove si colloca il maggior esponente in materia di process mining e confronto con l'attività svolta presso altri gruppi di ricerca con obiettivi affini;

Attività 3B: raffinamento dei metodi e realizzazione di caratteristiche specifiche per il dominio applicativo scelto;

Attività 3C: analisi dei risultati sperimentali ottenuti sul particolare dominio applicativo scelto;

Attività 3D: stesura della tesi di dottorato.

10) Valutazione dei risultati.

I risultati ottenuti dai vari prototipi sviluppati durante l'intero progetto di ricerca saranno valutati attraverso le modalità più adatte all'ambito applicativo in cui i modelli saranno adottati. Verranno condotte sperimentazioni seguendo protocolli formali che assicurano la replicabilità degli esperimenti oltre a stabilire la valenza scientifica dei modelli proposti.
In particolare, per i modelli di processo, ci sono quattro criteri di qualità che possono essere presi in considerazione: (a) fitness, (b) simplicity, (c) precision, e (d) generalization.

- Un modello con un buon fitness supporta la maggior parte dei comportamenti rilevati nei log (un modello ha fitness massimo se tutte le istanze nel log possono essere riprodotte sul modello dall'inizio alla fine).
- In accordo con il noto principio del Rasoio di Occam e il principio del Minimum Description Length, il modello migliore per descrivere il comportamento rilevato in un log è quello più semplice.
- Un modello è preciso se non ammette comportamenti che si discostano molto dai comportamenti riscontrati nel log. Un modello che non è preciso è "underfitting" (il modello sovrageneralizza il comportamento usuale nel log).
- Un modello dovrebbe anche generalizzare senza limitarsi solo ai comportamenti memorizzati nei log. Un modello che non generalizza è "overfitting" (indica che il modello generato è estremamente specifico).

11) Eventuali referenti esterni al Dipartimento

Durante gli studi e l'attività lavorativa, saranno selezionati alcuni referenti stranieri, operanti presso Università della Comunità Europea, al fine di supportare il lavoro di stesura della tesi di dottorato. Un possibile referente esterno è:
Wil van der Aalst - Department of Mathematics Computer Science of the Technische Universiteit Eindhoven