Big data and IoT for precision farming

Proposta di progetto

Le attività si svolgeranno nell'ambito del progetto WeLaser [1]. A seguito della pervasività di sistemi IoT e dell'avanzamento tecnologico, l'agricoltura sta vivendo un forte processo d'innovazione che coinvolge sistemi robotici, algoritmi di *pattern recognition*, e sistemi di predizione per la gestione di risorse e trattamenti agricoli. Solo in Europa vengono utilizzati circa 130 milioni di tonnellate di erbicidi all'anno che persistono nell'ambiente, distruggono piante e insetti benefici per il suolo e producono effetti sulla salute degli animali e degli esseri umani. La sostituzione degli erbicidi con sistemi robotici automatici è in fase di studio [1]. Tale innovazione si nutre di dati eterogenei provenienti da fonti diverse (ad esempio dati sensoriali, meteo e satellitari), il cui volume raddoppia ogni 2 anni [2]. I sistemi convenzionali di gestione e analisi dati convenzionali sono obsoleti, e lasciano spazio a sistemi big data e allo sviluppo d nuove tecniche analitiche. In questo contesto si aprono numerosi scenari di ricerca, tra cui il supporto alla gestione e all'analisi di dati legati all'agricoltura di precisione, sui quali si sono focalizzate precedenti attività di ricerca (AgroBigData science [3], Morefarming [4]). Le attività di seguito indicate sono proposte come prosecuzione dei precedenti lavori, allo scopo di integrare i risultati già ottenuti con nuovi contributi di ricerca. In particolare, gli obbiettivi sono (i) l'organizzare e orchestrare un *data lake*, ovvero repository di dati eterogenei provenienti da sorgenti diverse (ad esempio da sensori e sistemi IoT legati all'agricoltura di precisione), (ii) l'abilitare e il supportare l'analisi di tali dati per *data scientist* non necessariamente esperti di architetture big data e framework informatici.

Gestione di data lake per l'agricoltura di precisione

Introduzione. Oggigiorno, diverse tecnologie offrono soluzioni per la gestione di grossi volumi di dati prodotti con velocità in crescita costante. L'integrazione di dati eterogenei rimane un problema non risolto. Tale tematica è di rilievo e di oggetto di ricerca in ambito data lake, ovvero un repository---su piattaforma big data---che consente di archiviare grandi quantità di dati in formato nativo da cui un data scientist estrapola flussi di informazione tipicamente legati a sistemi di supporto alle decisioni. In un data lake, la varietà si manifesta nativamente, data l'inclusione di dati con formato, schema, e stati di elaborazione diversi (ad esempio, raw, parzialmente processato, e pronto per l'analisi).

Attività precedenti. Nelle precedenti attività di ricerca, sono state presentate tecniche per la profilazione di collezioni di dati (ossia individuare delle regole che associno i valori dei campi di un'istanza ad un determinato schema), con l'obiettivo di comprendere le logiche per cui esistono schemi diversi in una stessa collezione [5]. Inoltre, è stato proposto un approccio per abilitare le interrogazioni OLAP in un contesto *variety-aware* [6] e poliglotta, con l'idea di sfruttare l'espressività e l'eterogeneità degli schemi come valore aggiunto, sia in fase di interrogazione che di interpretazione del risultato.

Attività proposte. La gestione di robot in tempo reale richiede la creazione di servizi finalizzati alla raccolta/memorizzazione/analisi di dati IoT e di traiettoria. L'assenza di una gestione consistente e scalabile trasforma il data lake in una *swamp (palude)*, dove lo sforzo per la gestione e ricerca di informazione diventa più complessa dell'analisi stessa. L'obiettivo di ricerca è quello di sviluppare un *intelligent data lake a* supporto di decisioni basate su dati di movimento e sensoriali provenienti da robot e sistemi IoT [7], ovvero solidificare e integrare le precedenti attività di ricerca in un sistema capace di integrarsi in un data lake esistente e in grado di fornire strumenti a supporto di *data provenance* (cioè i metadati riguardo il processo di trasformazione di un *data item*) e *orchestration* (cioè le funzionalità legate ai flussi di dati). Da un lato, questo richiede (in collaborazione con gli esperti del settore agricolo) di progettare e implementare le funzioni di estrazione dal data lake della conoscenza necessaria a supporto

delle decisioni. Dall'altro lato, è necessario studiare lo stato dell'arte rispetto alle tecniche e tecnologie disponibili in contesto big data, al fine di adottare soluzioni all'avanguardia.

Analisi di dati di agricoltura di precisione

Introduzione. Il concetto di *data lake* è stato introdotto per indicare un repository in cui far confluire dati eterogenei provenienti da molteplici fonti IoT e a cui poter attingere per effettuare analisi di vario tipo. Le sfide da affrontare in questo contesto sono legate all'adozione delle metodologie e delle tecniche quanto più automatiche e automatizzabili per l'estrazione estrarre conoscenza dal data lake in maniera efficiente.

Attività precedenti. Nella precedente attività di ricerca si è lavorato su MoReFarming [4], progetto finalizzato a supportare il processo decisionale nell'ambito dell'agricoltura di precisione. I dati grezzi vengono ottenuti da diverse fonti e in diverse modalità e formati: immagini satellitari dall'Agenzia Spaziale Europea, dati meteo dall'ARPAE, dati sulle colture dai consorzi regionali, confini amministrativi dall'ISTAT, dati puntuali raccolti in-situ da appositi sensori. In questo contesto è stata definita un'architettura adeguata a gestire il carico di lavoro e si è impostata una complessa attività di estrazione, pulizia ed integrazione per rendere i dati fruibili ad un livello analitico.

Attività proposte. L'obiettivo di ricerca è quello di, sulla base dell'architettura esistente, sviluppare servizi di *monitoring & alerting* basati su tecniche analitiche e di machine learning su scala big data. Ad esempio, la valutazione delle performance dei robot e l'adattamento dei modelli di movimento ed estirpazione delle erbacce. Inoltre, il sistema si occupa di segnalare i robot con funzionamenti anomali. Da un lato, questo richiede (in collaborazione con gli esperti del settore agricolo) di progettare e implementare le funzioni di estrazione dal data lake della conoscenza necessaria a fornire il supporto di monitoraggio. Dall'altro lato, è necessario sviluppare sistemi in grado di imparare e adattare modelli e tecniche di analisi in real time.

Attività di progetto

Nell'ambito dell'anno di lavoro, l'assegnista sarà coinvolto nelle seguenti attività:

1. Nel contesto di *data lake*:
    a. Aggiornamento sulle soluzioni ad oggi proposte in letteratura
    b. Definizione delle modalità a supporto di un intelligent data lake
    c. Valutazione sperimentale delle performance e della scalabilità dell'acquisizione e integrazione di dati IoT e di mobilità.
2. Nel contesto di analisi su *big data*:
    a. Valutazione delle tecniche e delle tecnologie disponibili in letteratura
    b. Progettazione ed implementazione di funzioni di analytics per le performance dei robot
    c. Definizione di tecniche per individuare le configurazioni minime di sensori IoT necessari a supporto delle tecniche analitiche

Bibliografia

[1] WeLaser, 2020. H2020 "Sustainable weed management in agriculture with laser-based autonomous tool" — n. Grant Agreement 101000256 e n. Cup J32F20001250006.

[2] SINTEF. "Big Data, for better or worse: 90% of world's data generated over last two years." ScienceDaily. www.sciencedaily.com/releases/2013/05/130522085217.htm (accessed June 6, 2018).

[3] http://agrobigdatascience.it/

[4] http://www.morefarming.it/

[5] Enrico Gallinucci, Matteo Golfarelli, and Stefano Rizzi. Schema Profiling of Document Stores. In Proc. SEBD. Squillace Lido, Italy (2017), 1–8.

[6] Enrico Gallinucci, Matteo Golfarelli, and Stefano Rizzi. Variety-Aware OLAP of Document-Oriented Databases. In Proc. DOLAP. Wien, Austria (2018).

[7] FIWARE. https://www.fiware.org/