

Università degli Studi di Salerno

Dipartimento di Informatica

Dottorato di Ricerca in Informatica
Curriculum Internet of Things and Smart Technologies
XXXVIII Ciclo

Tesi di Dottorato / Ph.D. Thesis

Metadata-Driven Data Management and Analysis: From Profiling to Applications

Roberto STANZIONE

Supervisor: **Loredana CARUCCIO**

PhD Program Director: **Prof. Andrea DE LUCIA**

A.A 2024/2025

ABSTRACT

The digitalization of the public sector has led to the availability of large and heterogeneous collections of data, emphasizing the need for effective methods to extract, interpret, and exploit data-driven knowledge. To this end, Data Profiling techniques enable to automatically infer metadata that describe the structure, patterns, and relationships within datasets. Among these, data dependencies represent an increasingly exploited form of metadata that can be leveraged to support a variety of data management and analytical tasks. This thesis investigates the use data dependencies to design metadata-driven approaches for three relevant tasks: *Data Imputation*, *Concept Drift Detection*, and *Query Relaxation*. It also contributes to the advancement of data profiling itself, by proposing new methods for discovering, interpreting, and explaining dependencies. The **first part of the thesis** provides a comprehensive overview of data profiling techniques and introduces two novel proposals: S-DIMε, a semantic-aware dependency discovery algorithm capable of handling long textual attributes, and RYAN, a visualization and explanation tool that employs Large Language Models to describe the evolution of discovered dependencies over time. The **second part of the thesis** focuses on data quality, with particular attention to the imputation of missing data. It presents RIFT, a metadata-driven framework for imputing multivariate time series, and an empirical study demonstrating the mutual benefits of integrating data profiling and metadata-driven imputation in dynamic scenarios. The **third part of the thesis** explores additional use cases of metadata. In particular, it will introduce a novel metadata-driven method for concept drift detection, showing that monitoring the evolution of dependencies provides accurate drift estimations in machine learning pipelines. Furthermore, a metadata-driven query relaxation tool, namely REQUIRED, is introduced to support flexible and context-aware data exploration. Together, the promising results observed in the proposed contributions pave the way for future investigations aimed at developing more ro-

bust, interpretable, and metadata-centric paradigms that underlie increasingly complex data-driven tasks.

ABSTRACT (ITALIAN)

La digitalizzazione del settore pubblico ha portato alla disponibilità di collezioni di dati ampie ed eterogenee, evidenziando la necessità di metodi efficaci per estrarre, interpretare e sfruttare conoscenza dai dati. A tal fine, le tecniche di *Data Profiling* consentono di inferire automaticamente metadati che descrivono la struttura, i pattern e le relazioni all'interno dei dataset. Tra questi, le dipendenze tra dati rappresentano una forma di metadato sempre più utilizzata, che può essere sfruttata per supportare una varietà di attività di gestione e analisi dei dati. Questa tesi esplora l'uso delle dipendenze tra dati per lo sviluppo di approcci *metadata-driven* rivolti a tre compiti rilevanti: *Data Imputation*, *Concept Drift Detection* e *Query Relaxation*. Essa contribuisce inoltre all'avanzamento del campo del *Data Profiling*, proponendo nuovi metodi per il discovery, l'interpretazione e la spiegazione delle dipendenze. La **prima parte della tesi** fornisce una panoramica completa delle tecniche di *Data Profiling* e introduce due proposte innovative: S-DiMε, un algoritmo di discovery di dipendenze semantic-aware, capace di gestire attributi testuali di grandi dimensioni, e RYAN, uno strumento di visualizzazione e spiegazione che impiega *Large Language Model* per descrivere l'evoluzione delle dipendenze scoperte nel tempo. La **seconda parte della tesi** si concentra sulla qualità dei dati, con particolare attenzione all'imputazione dei valori mancanti. Essa presenta RIFT, un framework *metadata-driven* per l'imputazione di serie temporali multivariate, e uno studio empirico che dimostra i benefici reciproci derivanti dall'integrazione tra *Data Profiling* e imputazione guidata dai metadati in scenari dinamici. La **terza parte della tesi** esplora ulteriori casi d'uso dei metadati. In particolare, introduce un nuovo metodo *metadata-driven* per la rilevazione del *concept drift*, mostrando come il monitoraggio dell'evoluzione delle dipendenze consenta di stimare in modo accurato i cambiamenti nei modelli di apprendimento automatico. Inoltre, viene presentato uno strumento interattivo di *Query Relaxation*, denominato REQUIRED, che supporta un'esplorazione

dei dati flessibile e sensibile al contesto. Complessivamente, i risultati promettenti osservati nei contributi proposti aprono la strada a future ricerche volte a sviluppare paradigmi più robusti, interpretabili e incentrati sui metadati, che si pongono alla base di task guidati dai dati sempre più complessi.