ABSTRACT

Data Profiling represents one of the most crucial processes in data quality assessment. It includes a set of activities to efficiently analyze datasets and provide insights from them. Such activities rely on the identification of metadata to capture semantic relationships within data, and can be exploited for several purposes, such as optimizing queries, cleaning data, evaluating feasibility of machine learning models, and so forth. The types of metadata range from simple counters of attribute values or null values, to complex integrity constraints, such as functional dependencies (FDs), relaxed functional dependencies (RFDs), and inclusion dependencies (INDs). However, the discovery of these metadata represents an important challenge for data profiling tasks, since the number of possible metadata can be exponential with respect to the number of attributes, and requires analyzing a huge number of attribute combinations. To this end, several discovery algorithms have been proposed in the literature, with the aim of providing solutions in which the complexity of the search space is reduced by exploiting some theoretical properties of the different types of metadata. Although some of the discovery algorithms described in the literature achieve good performances, most of them are not suitable in dynamic scenarios, in which new data are frequently added and updated into the datasets. This need is widely growning with the proliferation of the Internet of Things (IoT) technologies, since it is necessary to define new algorithms capable of dynamically analyzing the streams of data they produce.

In this scenario, after reviewing basic data profiling tasks and applications, as well as basic notations for representing profiling metadata, this thesis starts presenting an innovative tool that extracts metadata from unstructured web data sources, aiming to derive a focused crawler. Then, the thesis focuses on the discovery problem of FDs and RFDs in static and dynamic scenarios, by analyzing their complexities and by introducing several new incremental methodologies and algorithms for discovering

FDs and RFDs, aiming to avoid the re-execution of the discovery process from scratch upon update operations on datasets. In particular, a first proposal is an evolutionary discovery algorithm for hybrid RFDs named REDEVO (RElaxed fD EVOlutionary discovery algorithm), which uses naturally inspired operations to iteratively browse candidates over the search space, few of which survive the evolution process. It identifies a broad class of RFDs, by evaluating each candidate by means of support and confidence quality measures as a fitness function.

Then, this thesis presents four new discovery algorithms for FDs and RFDs in dynamic scenarios. The first algorithm is named INCREMENTAL-FD, which is able to update the set of holding FDs upon insertions of new tuples to the data instance, without having to restart the discovery process from scratch. It exploits a bit-vector representation of FDs, and an upward/downward search strategy, aiming to reduce the overall search space. The algorithm represents the baseline for the definition of a second incremental algorithm for discovering FDs, named REXY (RegEX-based incremental discoverY). The latter adopts a new validation method that exploits Regular Expressions (RegExs) to improve the validation process for each FD candidate, by restricting the search to a subset of data. The third algorithm is named COD3, an efficient and incremental algorithm for discovering FDs holding on data streams. To the best of our knowledge, COD3 represents the first proposal to use a non-blocking architectural model to face the problem of FD discovery from data streams. It relies on a novel data structure, named Validation Graph, which enables a fast exploration of the search space according to the discovered FDs, leading to a new fast FD validation process. The last algorithm presented in this thesis is BIRD (Bit-vector based Incremental RFD_e Discoverer), which tackles the problem of the incremental discovery of RFDs. BIRD analyzes how new inserted tuples impact on candidate RFDs, checking whether they invalidate some previously holding ones, and possibly generating new candidates. Moreover, BIRD is able to split the discovery process into level-wise parallel executions.

Finally, the thesis describes three new tools designed and developed for monitoring incremental discovery algorithms during their executions. These tools enable users to properly visualize the trend of FDs and

RFDs discovered over time, provide an overview at each time instant of the correlation between attributes included in the discovery results, compare FDs and RFDs resulting from different executions of the discovery algorithms, and directly manipulate discovery results through visual metaphors.

ABSTRACT

I processi di Data Profiling rappresentano uno strumento chiave per supportare la valutazione della qualità dei dati. Questi si articolano in attività rivolte all'analisi efficiente di insiemi di dati al fine di estrarre informazioni utili dagli stessi. Tali attività permettono di catturare relazioni semantiche all'interno dei dati attraverso l'estrazione di metadati, i quali possono essere sfruttati per diversi scopi, come l'ottimizzazione delle query, la pulizia dei dati e la valutazione dei modelli di machine learning. I tipi di metadati possono variare da semplici contatori sul numero di valori nulli o di valori distinti in un dataset, a vincoli di integrità, come dipendenze funzionali (Functional Dependency - FD), dipendenze funzionali rilassate (Relaxed Functional Dependency - RFD) e dipendenze di inclusione (Inclusion Dependency - IND). Tuttavia, la complessità del problema di discovery automatico di questi metadati rappresenta una delle principali sfide per le attività di data profiling, poiché il numero di possibili metadati potrebbe essere esponenziale rispetto al numero di attributi dei dataset, considerando la quantità di combinazioni di attributi da analizzare. A tal fine, sono stati proposti in letteratura diversi algoritmi per il loro discovery automatico dai dati, con l'obiettivo di fornire soluzioni in cui la complessità dello spazio di ricerca viene ridotta sfruttando alcune proprietà teoriche dei diversi tipi di metadati. Sebbene alcuni di questi algoritmi di discovery raggiungano buone prestazioni, la maggior parte di essi non è in grado di effettuare processi di discovery automatico in scenari dinamici, in cui si considera la possibilità che i dati vengano frequentemente aggiornati. Questa esigenza è ampiamente cresciuta con la diffusione dell'Internet of Things (IoT), poiché gli algoritmi dovrebbero essere in grado di analizzare dinamicamente i flussi di dati che questi strumenti intelligenti producono.

In questo scenario, dopo aver esaminato le attività e le applicazioni del data profiling, e dopo aver introdotto le notazioni di base per la rappresentazione dei metadati, questa tesi presenta uno strumento innovativo

che estrae i metadati da sorgenti di dati Web non strutturati, con l'obiettivo di derivare un crawler mirato. Successivamente, la tesi si concentra sul problema del discovery di FD e RFD in scenari statici e dinamici, analizzando la complessità del problema di discovery e introducendo diverse nuove metodologie e algoritmi incrementali per l'estrazione automatica di FD e RFD, con l'obiettivo di evitare la riesecuzione del processo di discovery dall'inizio dopo le operazioni di aggiornamento dei dati. In particolare, una prima proposta presentata è un algoritmo di discovery evoluzionistico per RFD ibride chiamato REDEVO (RElaxed fD EVOlutionary discovery algorithm), che utilizza operazioni ispirate alla selezione naturale delle specie per analizzare iterativamente insiemi di dipendenze candidate che evolvono, poche delle quali sopravviveranno al processo di evoluzione. REDEVO permette di identificare un'ampia classe di RFD, effettuando la validazione di ciascuna candidata mediante le misure di support e confidence, le quali definiscono la funzione di fitness.

Successivamente, questa tesi presenta quattro nuovi algoritmi di discovery di FD e RFD in scenari dinamici. Il primo algoritmo è denominato INCREMENTAL-FD, il quale è in grado di aggiornare l'insieme di FD valide dopo l'inserimento di nuove tuple nei dati, senza dover rieseguire completamente il processo di discovey. L'algoritmo sfrutta una rappresentazione vettoriale binaria delle FD e una strategia di ricerca in discesa/salita, con l'obiettivo di ridurre lo spazio di ricerca da analizzare. L'algoritmo INCREMENTAL-FD rappresenta la baseline per la definizione di un secondo algoritmo incrementale per il discovery di FD, chiamato REXY (RegEX-based incremental discoverY). Quest'ultimo adotta un nuovo metodo di validazione che sfrutta le espressioni regolari (RegEx) per migliorare il processo di validazione di ogni FD candidata, limitando la validazione al sottoinsieme di dati interessato dalle modifiche. Il terzo algoritmo incrementale, denominato COD3, permette di effettuare il discovery di FD su data stream. Al meglio della nostra conoscenza, COD3 rappresenta la prima proposta in letteratura che utilizza un modello architetturale non bloccante per affrontare il problema di discovery di FD dai flussi di dati. Esso pone le sua fondamenta su una nuova struttura dati, denominata Validation Graph, che consente una rapida esplorazione dello spazio di ricerca in base alle FD precedentemente inferite dai dati, e attaver-

so la quale è stato possibile definire un nuovo ed efficente processo di validazione di FD. L'ultimo algoritmo presentato in questa tesi è BIRD (Bit-vector based Incremental RFD_e Discoverer), che affronta il problema del discovery incrementale di RFD. BIRD analizza come le nuove tuple inserite impattano sulle RFD, verificando se tali tuple comportano l'invalidazione di alcune RFD già presenti ed eventualmente generando nuove RFD candidate. Inoltre, BIRD è in grado di suddividere il processo di discovery in esecuzioni parallele per ogni livello dello spazio di ricerca.

Infine, la tesi presenta tre nuovi tool progettati e sviluppati per monitorare i processi di discovery incrementali durante l'esecuzione degli algoritmi. Questi strumenti consentono agli utenti di visualizzare continuamente l'andamento e l'evoluzione delle FD e RFD estratte dai dati nel corso del tempo, di analizzare in ogni istante temporale la correlazione tra gli attributi inclusi nei risultati di discovery, e di confrontare FD e RFD estratte in diverse esecuzioni degli algoritmi e di manipolarle mediante nuove metafore visuali.