Università degli Studi di Salerno
Dipartimento di Scienze Aziendali - Management & Innovation Systems

Dottorato di Ricerca in Management and Information Technology
Curriculum Informatica, Sistemi Informativi e Tecnologie del Software
XVI ciclo

Tesi di dottorato in

Relaxed Functional Dependencies: Definition, Discovery and Applications

Anno Accademico 2016-2017

Candidato
Loredana Caruccio

Coordinatore
Prof. Andrea De Lucia

Tutor
Prof. Giuseppe Polese

Co-Tutor
Prof. Vincenzo Deufemia

Abstract

Functional dependencies (FDs) were conceived in the early '70s, and were mainly used to verify database design and assess data quality. However, to solve several issues in emerging application domains, such as the identification of data inconsistencies, patterns of semantically related data, query rewriting, and so forth, it has been necessary to extend the FD definition. This has lead to new definitions of functional dependencies, named Relaxed Functional Dependencies (RFDs), since they relax some of the constraints of the FD definition. Based on the types of constraints they relax RFDs can be classified into three categories. In particular, there exist RFDs relaxing on the data comparison, i.e., by considering data similarity rather than equality, those relaxing on the extent, i.e., by admitting the possibility that the RFD holds on a subset of data, and finally, those relaxing on both criteria. Thresholds might be used in all such categories, either to specify the similarity degree or to indicate the percentage of validity with respect to the entire database.

Moreover, while FDs were originally specified at database design time as properties of a schema that should hold on every legal instance of it, to effectively employ them in emerging application domains it became necessary to devise techniques to automatically discover them from big data collections. However, while in the literature, there are several discovery algorithms for FDs, few RFD definitions are equipped with algorithms for detecting them from data.

In this scenario, this thesis introduces the general definition of Relaxed Functional Dependency, and characterizes the state of the art of RFDs, providing a classification criteria, motivating examples, and a systematic analysis of them. Additionally, it describes several techniques and algorithms to discover them from data. In particular, a first proposal,

named DiMε (Difference Matrices and ε-thresholds), represents a generic method capable of detecting and validating any type of candidate RFD relaxing on the data comparison method and/or the extent, according to per-specified thresholds. It starts from user-specified thresholds, and performs a level-by-level generation of candidate dependencies to be successively validated. Moreover, in DiMε several pruning techniques have been proposed in order to reduce the search space during the discovery process. Forasmuch as the discovery problem for RFDs is a complex one, even when thresholds are given in input, a genetic algorithm, named GA-RFD, has been proposed. It identifies a broad class of RFDs, including both those relaxing on the comparison method and those relaxing on the extent. In particular, GA-RFD uses operations inspired to natural specie evolutions (such as, natural selection, crossover, and mutation) to iteratively generate new candidate RFDs, few of which survive to the evolution process. The selection of candidates that must survive is accomplished by means of a fitness function, which exploits the support and confidence quality measures.

The last proposal for the discovery of RFDs from data presented in this thesis is the algorithm DOMINO, which exploits the concept of dominance, from the multi-attribute utility theory, to detect RFDs relaxing on the attribute comparison. The strength point of DOMINO is that is able to determine the features of RFDs without requesting the user to specify input parameters.

Finally, RFDs have been applied to many application contexts, and can be potentially applied into many additional domains. Indeed, many new RFD definitions have been introduced to solve problems in specific application domains, such as the matching dependencies for the record matching problem, or the conditional functional dependencies for the data cleaning. The last part of this thesis describes additional potential application case studies for RFDs such as query relaxation, query/view synchronization, and data integration. In particular, RFDs can be employed in such domains to highlight relationships among data that can facilitate the detection of potential solutions.

Abstract

Le dipendenze funzionali (FD) sono state introdotte agli inizi degli anni '70, quando venivano maggiormente utilizzate per valutare la progettazione dei database e garantire la qualità dei dati. Tuttavia, si è avuta la necessità di estendere la definizione di dipendenza funzionale allo scopo di risolvere diverse problematiche in domini applicativi emergenti, tra cui l'identificazione delle inconsistenze nei dati e/o di pattern di dati semanticamente correlati, la necessità di effettuare riscritture delle query, e così via. Questa necessità ha portato alla nascita di nuove definizioni di dipendenze funzionali, che sono state chiamate dipendenze funzionali rilassate (RFD), in quanto queste rilassano alcuni dei vincoli della definizione canonica di FD. Sulla base del tipo del vincolo che rilassano le RFD possono essere classificate in tre categorie. In particolare, esistono RFD che rilassano sul confronto dei dati, ovvero considerando la similarità tra i dati piuttosto che l'uquaglianza; altre RFD, invece, rilassano sul grado di soddisfacilità, denominato extent, in quanto permettono di avere la possibilità che la RFD valga per un sottoinsieme dei dati e, infine, dipendenze che rilassano su entrambi i criteri. Inoltre, in genere per poter specificare il grado di somiglianza ammesso o indicare la percentuale di valità rispetto alla quantità totale di dati presenti nel database, vengono utilizzate delle soglie.

Inoltre, mentre le FD venivano originariamente specificate durante la progettazione dei database come proprietà di uno schema che dovrebbero essere soddisfatte da ogni istanza ammissibile del database stesso, è diventato necessario definire tecniche per la scoperta automatica delle dipendenze partendo da grosse collezioni di dati allo scopo di utilizzarle nei domini applicativi emergenti. Tuttavia, sebbene in letteratura esistono diversi algoritmi di scoperta delle dipendenze per le FD, soltanto poche delle nuove definizioni di RFD sono

corredate da algoritmi per la loro scoperta dai dati.

In questo scenario, questo lavoro di tesi introduce la definizione generale di dipendenza funzionale rilassata, e caratterizza lo stato dell'arte delle RFD, fornendo criteri di classificazione, esempi movitazionali, e un'analisi sistematica di tali dipendenze. Inoltre, esso descrive diverse tecniche ed algoritmi per la scoperta delle dipendenze dai dati. In particolare, una prima proposta, denominata DiMε (Difference Matrices and ε-thresholds), rappresenta una metodologia generale capace di scoprire e validare qualsiasi tipo di RFD che applica un rilassamento sul metodo di confronto dei dati e/o sull'extent, in accordo a soglie predefinite. Infatti, questo partendo da soglie specificate dall'utente, effettua una generazione di RFD candidate per livelli, le quali vengono successivamente validate. Inoltre, in DiMε sono state proposte diverse tecniche di taglio allo scopo di ridurre lo spazio di ricerca durante il processo di scoperta. In generale, la scoperta delle RFD dai dati è un problema complesso, anche quando vengono fornite in input le soglie. Per tale motivo, è stato proposto un algoritmo genetico per affrontare tale problema, denominato GA-RFD. Questo permette l'identificazione di un'ampia classe di RFD, ovvero sia RFD che rilassano sul metodo di confronto che quelle che rilassano sull'extent. In particolare, GA-RFD usa operazioni ispirate all'evoluzione delle specie naturali (tali come, la selezione naturale, il crossover, e la mutazione), allo scopo di generare nuove RFD candidate, poche delle quali sopravvivono al processo di evoluzione. La selezione delle RFD candidate viene realizzata attraverso una funzione di fitness che sfrutta le misure di qualità di supporto e confidenza.

L'ultima proposta sulla scoperta delle RFD dai dati presentata in questa tesi è l'algoritmo DOMINO, il quale sfrutta il concetto di dominanza, che ha origine nella teoria dell'utilità multi-attributo, allo scopo di individuare le RFD senza richiedere all'utente di specificare parametri in input.

Infine, le RFD sono state utilizzate in molti contesti applicativi, e possono essere potenzialmente applicate in molti altri domini. Infatti, molte delle nuove definizioni di RFD sono state introdotte per risolvere problemi di specifici domini applicativi, come ad esempio le matching dependency per il problema del record matching, o le conditional functional dependency per il data cleaning. L'ultima parte di questa tesi descrive ulteriori casi di studio di potenziali applicazioni per le RFD, ovvero il query relaxation e la sincronizzazione

delle query e delle viste. In particolare, le RFD possono essere applicate in questi domini per poter evidenziare le relazioni tra i dati e facilitare l'applicazione delle potenziali soluzioni.