



Università degli Studi di Salerno

Dipartimento di Informatica

---

Tesi di Laurea Triennale in

Informatica

**UN ALGORITMO PER L'INFERENZA DI  
DIPENDENZE FUNZIONALI RILASSATE:  
IDENTIFICAZIONE DI SOGLIE OTTIME**

**Relatore**

Chiar.mo Prof. Vincenzo Deufemia

**Secondo Relatore**

Dott.ssa Loredana Caruccio

**Candidato**

Giovanni Leo

**Matr.** 0512103062

---

Anno Accademico 2016-2017



*Alla mia famiglia, per aver creduto in me e per avermi sostenuto.*

*Alla mia ragazza, per avermi supportato e sopportato.*

# Abstract

Nella progettazione di una base di dati ci sono aspetti essenziali da prendere in considerazione per assicurare un servizio quanto più efficiente possibile. Considerato il netto aumento del flusso di dati degli ultimi anni, la *data quality* è divenuta una materia estremamente interessante vista la cospicua presenza di dati "sporchi" nelle basi di dati. Per ridurre anomalie ed inconsistenze ci vengono incontro le *Dipendenze funzionali*, utilizzate ampiamente per stabilire vincoli di integrità tra i dati. La grande mole di dati, però, ha reso necessario un riadattamento delle dipendenze funzionali rendendole in grado di catturare inconsistenze più ampie nei dati. Le *Dipendenze funzionali rilassate o approssimate* (**RFD**) sono da considerarsi come una naturale evoluzione o generalizzazione delle *dipendenze funzionali canoniche*. Il concetto più importante introdotto dalle RFD è quello della *similarità*. Nelle dipendenze funzionali classiche esisteva soltanto il concetto di uguaglianza tra dati, nelle RFD espandiamo questo concetto ad una similarità, questo ci permetterà di coprire una quantità di dati maggiore. Tuttavia le RFD possono fornire vantaggi solo se possono essere scoperte automaticamente dai dati. Il lavoro di tesi si è basato su questo ultimo concetto di ottenere le RFD in seguito

ad una procedura automatizzata. Durante le varie fasi di studio si è pensato ed implementato un algoritmo che permette, attraverso tre fasi intermedie, la scoperta di RFD di un dataset dato come input. Per questo lavoro di tesi mostreremo l'idea dell'algoritmo generale ed entreremo nel dettaglio dell'ultima fase di sviluppo(RFD Discovery), mostrando, infine, i risultati della sperimentazione.

# Indice

<b>1</b>	<b>Introduzione</b>	<b>1</b>
1.1	Incipit . . . . .	1
1.2	Nozioni Preliminari . . . . .	3
1.3	Stato dell'Arte . . . . .	4
1.4	Studi Preliminari . . . . .	4

# Elenco delle tabelle

1.1	Esempio di schema di relazione . . . . .	3
-----	--	---

## Snippet di codice



# Introduzione

## 1.1 Incipit

Nella progettazione di una base di dati ci sono aspetti essenziali da prendere in considerazione per assicurare un servizio quanto più efficiente possibile. Uno di questi servizi è certamente la *qualità dei dati*, una base di dati con questa caratteristica farà sì che le inconsistenze tra i dati siano il minor numero possibile. Negli ultimi anni la crescita delle reti ha portato ad un aumento considerevole del flusso di dati rendendo la *data quality* una materia estremamente interessante vista la cospicua presenza di dati "sporchi" proveniente da fonti differenti. Per ridurre questo tipo di anomalie è impensabile tentare di eliminare le *inconsistenze* manualmente, una procedura di questo tipo può essere facilmente incline ad errori soprattutto con la quantità di dati precedentemente citata. In questo lavoro ci vengono incontro le *Dipendenze funzionali*, utilizzate ampiamente per stabilire vincoli di integrità tra i dati e ridurre anomalie e inconsistenze all'interno della nostra base di dati. La grande mole

di dati, però, ha reso necessario un riadattamento delle dipendenze funzionali rendendole in grado di catturare inconsistenze più ampie nei dati. Le *Dipendenze funzionali rilassate o approssimate (RFD)* sono da considerarsi come una naturale evoluzione o generalizzazione delle *dipendenze funzionali canoniche*. Questo nuovo strumento ci permette di adattare le semplici dipendenze funzionali a diversi contesti applicativi, infatti, le RFD possono applicarsi anche solo ad una porzione di database. Il concetto più importante introdotto dalle RFD è quello della *similarità*. Nelle dipendenze funzionali classiche esisteva soltanto il concetto di uguaglianza tra dati, nelle RFD espandiamo questo concetto ad una similarità, questo ci permetterà di coprire una quantità di dati maggiore. Tuttavia le RFD possono fornire vantaggi solo se possono essere scoperte automaticamente dai dati. Il lavoro di tesi si è basato su questo ultimo concetto di ottenere le RFD in seguito ad una procedura automatizzata. Durante le varie fasi di studio si è pensato ed implementato un algoritmo che permette, attraverso tre fasi intermedie, la scoperta di RFD di un dataset dato come input. Le tre fasi di questo algoritmo sono: *Feasibility, Minimality, RFD Discovery*. Per questo lavoro di tesi mostreremo l'idea dell'algoritmo generale ed entreremo nel dettaglio dell'ultima fase di sviluppo (RFD Discovery), mostrando, infine, i risultati della sperimentazione. Per questo algoritmo, particolare attenzione è stata posta sull'efficienza, oltre che sull'efficacia, studiando un'implementazione basata sul multithreading e predisponendola ad eventuale adattamento parallelo.

## 1.2 Nozioni Preliminari

Prima di esporre le RFD è necessario introdurre alcuni concetti preliminari.

**Schema di relazione** Uno schema di relazione è costituito da un simbolo  $R$ , detto nome della relazione, e da un insieme di attributi  $X = \{A_1, A_2, \dots, A_n\}$ , di solito indicato con  $R(X)$ . A ciascun attributo  $A \in X$  è associato un dominio  $dom(A)$ . Uno schema di base di dati è un insieme di schemi di relazione con nomi diversi:

$$R = \{R_1(X_1), R_2(X_2), \dots, R_n(X_n)\}.$$

Una relazione su uno schema  $R(X)$  è un insieme  $r$  di tuple su  $X$ . Per ogni istanza  $r \in R(X)$ , per ogni tupla  $t \in r$  e per ogni attributo  $A \in X$ ,  $t[A]$  rappresenta la proiezione di  $A$  su  $t$ . In modo analogo, dato un insieme di attributi  $Y \subseteq X$ ,  $t[Y]$  rappresenta la proiezione di  $Y$  su  $t$ . [1]

Matricola	Cognome	Nome	Data di nascita
123456	Rossi	Maria	25/11/1991
654321	Neri	Anna	23/04/1992
456321	Verdi	Fabio	12/02/1992

Tabella 1.1: Esempio di schema di relazione

### 1.3 Stato dell'Arte

### 1.4 Studi Preliminari

# Bibliografia

- [1] F. P. P. S. T. R. Atzeni P., Ceri S., *Basi di dati: Modelli e linguaggi di programmazione*. McGraw Hill, 2013.