Università degli Studi di Milano - Bicocca
Scuola di Scienze
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea Magistrale in Informatica

Tecniche di Record Linkage

Alberici Federico - 808058
Bettini Ivo Junior - 806878
Cocca Umberto - 807191
Traversa Silvia - 816435

Anno Accademico 2019 - 2020

Indice

Ricerca

Data Quality

La consapevolezza del peso che dati di alta qualità hanno nel supportare decisioni informate e, viceversa, delle conseguenze disastrose cui dati inaccurati possono portare, è cresciuta di pari passo con il diffondersi delle fonti informative a disposizione delle organizzazioni, creando sempre più forte l'esigenza di una gestione adeguata della qualità dei dati aziendali.
La ricerca sulla qualità dei dati è iniziata correttamente negli anni '90 e varie definizioni di ciò sono state definite nel corso degli anni.
Un gruppo di ricerca del MIT, guidato da il professor Wang, ha definito la qualità dei dati come condizione per l'uso e ha proposto il suo giudizio dipende dai suoi consumatori. Allo stesso tempo, hanno definito una "dimensione della qualità dei dati" come un insieme di attributi di qualità dei dati che rappresentano un singolo aspetto o costrutto della qualità dei dati.
Sono necessarie tecniche di misurazione completa per consentire alle organizzazioni di valutare lo stato della loro qualità delle informazioni organizzative e monitorarne il miglioramento.
Ma cosa si intende quando si parla di qualità dei dati e come si misura?
Le best practice in questo ambito suggeriscono l'utilizzo di opportune metriche per la definizione e la misurazione della qualità del dato. Tra le metriche più comuni troviamo:

- **completezza**, i dati raccolti bastano per rappresentare l'informazione necessaria;
- **accuratezza**, la precisone dei dati;
- **tempestività**, i tempi di acquisizione dei dati sono utili per il processo;
- **coerenza**, i dati non sono contradditori tra di loro;
- **univocità**, i dati rappresentativi della stessa informazione presenti in diversi componenti del sistema informativo assumono lo stesso valore;
- **integrità**, i dati presenti nel sistema informativo corrispondono a quelli originariamente immessi;

- **conformità formale**, i dati immessi nel sistema informativo rispettano gli standard formali appositamente definiti.

In tempi attuali, è emerso un altro tipo problema: i big data. Analisi e ricerca complete di standard di qualità e metodi di valutazione della qualità per questo tipo di informazioni attualmente manca o non è completa.

Metodoligia Data Quality

Il professor Batini definisce la metodologia di qualità dei dati come un insieme di linee guida e tecniche che, a partire dalle informazioni di input che descrivono un determinato contesto applicativo, ne deriva un processo razionale per valutare e migliorare la qualità dei dati. Ci sono tre fasi principali per tale attività:

- **ricostruzione dello stato**, al fine di ottenere due informazioni contestuali, facoltative se sono già disponibili per l'uso;
- **valutazione e misurazione**, misurazione della qualità lungo dimensioni della qualità pertinenti o valutazione, quando tali misurazioni vengono confrontate con i valori di riferimento;
- **miglioramento**, attività che mirano per raggiungere nuovi obiettivi di qualità dei dati.

Miglioramento

Il miglioramento della qualità dei dati può essere effettuato attraverso strategie basate sui dati o sui processi. Nel primo caso, le tecniche più diffuse sono quella di standardizzazione (o normalizzazione), il record linkage e l'integrazione degli schemi e dei dati, mentre nel secondo caso si adotta un processo di ricostruzione. Nel caso del nostro progetto, per poter migliorare la qualità del dato abbiamo deciso di utilizzare la tecnica del record linkage.

Standardizzazione

Questo processo, chiamato anche normalizzazione, sostituisce per esempio una diversa ortografia di una parola con una sola ortografia.

Comparazione stringhe

Gli errori tipografici rendono impossibile confrontare esattamente tra di loro le stinghe. Per poter fare ciò, quindi, serve una funzione che cerca di trovare un punto di accordo tra i dati. Ci sono stati diversi tentativi di fornire questa funzione:

- Jaro ha proposto un comparatore di stringhe che tiene conto di inserimenti, eliminazioni e trasposizioni necessarie per abbinare le due stringhe;
- Winkler ha proposto una variante della distanza Jaro (Jaro-Winkler);
- la distanza q-gram conta il numero di q caratteri consecutivi che concordano tra due corde;
- la distanza di edit classica, che conta il numero di operazioni (inserimenti, eliminazioni, modi cazioni) necessarie per abbinare le due stringhe

Record Linkage

Metodologie (?) di Record Linkage

Tools usati

Python Record Linkage Toolkit

Python Record Linkage Toolkit è una libreria che permette di effettuare record linkage sia in una sola fonte di dati che in multiple. Il package contiene metodi di indexing, come blocking e sorted neighbourhood indexing, funzioni per il confronto con diverse misure di similarità possibili e diversi algoritmi di classificazione, sia supervisionati che non.

Esperimenti

Dataset

Il dataset scelto per effettuare i nostri esperimenti contiene elenchi di ristoranti di Manhattan provenienti da 12 siti web, presi settimanalmente da Gennaio a Marzo 2009. Sono riportati in tutti il nome del ristorante, l'indirizzo e la città.

Pulizia dei dati

Prima di applicare le tecniche di record linkage è stata effettuata una pulizia dei dati. Avevamo a disposizione 7 file .txt, uno per ogni settimana considerata, contenente dati appartenenti a tutti i siti web. In totale erano presenti 215555 record, suddivisi come illustrato nella tabella:

file	**records**
restaurants_2009_1_22.txt	30401
restaurants_2009_1_29.txt	30775
restaurants_2009_2_05.txt	30805
restaurants_2009_2_12.txt	30863
restaurants_2009_2_19.txt	30876
restaurants_2009_2_26.txt	30898
restaurants_2009_3_12.txt	30937

Table 1: Record presenti nei file txt forniti

In particolare, per ogni ristorante è presente il seguente numero di record:

restaurant	records
ActiveDiner	6184
DiningGuide	814
FoodBuzz	2079
MenuPages	13143
NewYork	1774
NYMag	5124
NYTimes	3095
OpenTable	1539
SavoryCities	4536
TasteSpace	3635
TimeOut	14007
VillageVoice	2684

Table 2: Record per ristorante

Risultati

Record Linkage

Gold Standard