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 dal professor Wang, ha definito la qualità dei dati come condizione per il loro utilizzo e ha proposto il loro giudizio dipendentente dai consumatori finali. Allo stesso tempo, hanno definito una "dimensione della qualità dei dati" come un insieme di attributi 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 qualità delle informazioni organizzative e monitorarne il miglioramento.

Ma cosa si intende quando si parla di qualità dei dati e come si misura? La qualità dei dati è una caratteristica dei dati che ha a che fare con la loro abilità di soddisfare le esigenze e le aspettative implicite o esplicite dell'utente.

Per quanto riguarda la definizione di metrica ci riferiamo alle definizioni all'interno dello standard ISO 9126-1 e framework ISM3:

- una procedura (o metodo) di misurazione, cioè un algoritmo che prende l'elemento per misurare e lo associa a misura (sia esso valore ordinale o intervallo);
- una corretta unità di misura (o scala), ovvero di dominio di valori restituiti dalla procedura di misurazione. In generale, è possibile associare diverse metriche a cias-

cuna dimensione di qualità;

Questa può essere espressa attraverso molteplici dimensioni

Le best practice in questo ambito suggeriscono l'utilizzo di opportune metriche per la definizione e la misurazione. 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.

Il punto focale rimane sempre il dominio sul quale si vuole effettuare un processo di miglioramento dei dati. In base alla finalità delle informazioni e alle caratteritiche dei loro consumatori l'attenzione si rivolge ad un sottoinsieme di tale metriche.
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. Questo topic pone una serie di nuove sfide, dettate dalle caratteristiche intrinseche dei big data, riassumibili in quelle che sono chiamate "le 5 V":

- **Volume**: ingente massa di informazioni, in crescita vertiginosa, che non è possibile raccogliere con tecnologie tradizionali;
- **Velocity**: i dati nascono e vengono acquisiti sempre più rapidamente, con necessità di analisi in tempo reale;
- **Variety**: differenti tipologie di dati disponibili, provenienti da un numero crescente

di fonti eterogenee;

- **Veracity**: i dati devono essere affidabili, raccontare il vero;
- **Value**: abilità di trasformare una grande mole di dati in business.

Metodologia 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 [**10.1145/1541880.1541883**]. 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 [**census/jaro76**] nel 1976 ha proposto un comparatore di stringhe che tiene conto di inserimenti, eliminazioni e trasposizioni necessarie per abbinare le due stringhe;
- Winkler [**winkler90**] nel 1990 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

Il record linkage (conosciuto anche come data matching) è l'operazione che consiste nel trovare "records" che si riferiscono alla stessa entità, in dataset presi da differenti risorse (come ad esempio file, libri, siti e database). Questa operazione diventa necessaria quando vogliamo unire dei dataset differenti basati su dati simili che potrebbero avere o non avere lo stesso identificativo.
L'idea moderna di record linkage nasce alla fine degli anni cinquanta e viene formalizzata qualche tempo dopo da Ivan Fellegi e Alan Sunter [**fellegi69**] che, attraverso il loro lavoro, hanno dimostrato che le regole di decisione probabilistiche sono ottimali quando i dati che vengono confrontati sono condizionatamente indipendenti.
A partire dalla fine degli anni novanta, differenti tecniche di machine learning sono state sviluppate per poter capire, con condizioni favorevoli, la probabilità condizionata richiesta dalla teoria Fellegi-Sunter.
Il record linkage può essere interamente eseguito senza l'aiuto di un computer, ma il motivo principale per cui esso viene utilizzato è perchè si vuole ridurre o eliminare le

modifiche "fatte a mano" e per rendere più facile l'ottenimento del risultato. L'utilizzo del computer ha anche il vantaggio di utilizzare un processo di supervisione centrale, miglior qualità di controllo, velocità, consistenza e migliore riproducibilità dei risultati.

Metodologie di Record Linkage

Il record linkage si divide generalmente nei seguenti step:

1. vengono dati in input dei dataset;
2. viene definito uno spazio iniziale di ricerca;
3. si cerca di ridurre lo spazio di ricerca tramite un processo di "blocking" ;
4. viene definito lo spazio ridotto di ricerca;
5. vengono comparati i dati e viene presa una decisione;
6. vengono definitie le regole di matching, possibile matching o se i dati non sono collegabili;
7. secondo le regole definite è generato il dataset di output.

Data preprocessing Il record linkage è molto sensibile alla qualità dei dati che devono essere collegati, quindi idealmente prima di svolgere questa operazione ogni dataset deve essere controllato affinche la qualità sia delle migliori. Avvengono delle operazioni di standardizzazione, che consistono nel trasformare i dati o procedure più complesse, come ad esempio la tokenizzazione.

Entity resolution L'entity resolution è un processo di operazioni intelligenti che permettono alle organizzazioni di connettere i dati più disparati attraverso la possibilità di capire i matches tra le entità e le relazioni non ovvie fra i diversi dati.

Essa analizza tutte le informazioni collegate ad una entità prese da diverse sorgenti e cerca, attraverso un calcolo probabilistico, di determinare quali entità sono collegate e quali collegamenti (non ovvi) esistono fra loro.

Deterministic record linkage La metodologia più semplice di record linkage è chiamata "deterministica".

Essa genera collegamenti basati sul numero di singoli identificatori che hanno una

corrispondenza fra i dati dei dataset.

Due record si dicono *collegati* con una procedura di record linkage deterministico se se tutti o alcuni degli identificatori sono identici.

Questo metodo è una buona opzione se si stanno utilizzando dei dataset con delle entità che sono identificate da un id comune.

Probabilistic record linkage Il record linkage probabilistico, chiamato anche *fuzzy matching*, utilizza un approccio differente per poter collegare i dati. Viene tenuto conto di una gamma più ampia di potenziali identificatori, calcolando i pesi per ciascun identificatore in base alla sua capacità stimata di identificare correttamente una corrispondenza o una non corrispondenza. Questi pesi sono dunque usati per calcolare la probabilità che due dati registrati si riferiscano alla stessa entità.

Le coppie di record con probabilità al di sopra di una determinata soglia sono considerate corrispondenze, viceversa le altre sono considerate non corrispondenze. Le coppie che rientrano tra queste due soglie sono considerate "possibili corrispondenze" e possono essere trattate di conseguenza (ad esempio, revisioni umane, collegate o non collegate, a seconda dei requisiti). Mentre il collegamento deterministico dei record richiede una serie di regole potenzialmente complesse da programmare in anticipo, i metodi probabilistici di collegamento dei record possono essere "addestrati" per funzionare bene con un intervento molto meno umano.

Machine learning Negli ultimi anni, sono state utilizzate varie tecniche di apprendimento automatico per collegamento automatico. È stato riconosciuto che l'algoritmo classico per il collegamento probabilistico dei record sopra descritto è equivalente all'algoritmo Naive Bayes nel campo dell'apprendimento automatico, e utilizza la stessa assunzione dell'indipendenza delle sue caratteristiche (un presupposto che in genere non è vero).

È possibile ottenere una maggiore precisione utilizzando varie altre tecniche di apprendimento automatico, incluso un percettrone a strato singolo. Insieme alle tecnologie distribuite, l'accuratezza e la scala per il collegamento dei record possono essere ulteriormente migliorate.

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 su cui sono stati effettuati gli esperimenti è un elenco di ristoranti di Manhattan, estratti settimanalmente da Gennaio a Marzo 2009 da 12 siti web. L'unico attributo comune in tutti i dataset è il nome del ristorante, informazioni aggiuntive (come l'indirizzo o il quartiere) non sono presenti in modo uniforme.

Analisi dei dati

I dati sono messi a disposizione in sette file di testo, uno per ogni settimana considerata, ciascuno contenente informazioni appartenenti a tutti i siti web. In totale sono presenti 215555 record, suddivisi come illustrato nella seguente 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

E sono presenti le seguenti informaizoni:

- **ActiveDiner**: nome ristorante, indirizzo, paese
- **DiningGuide**
- **FoodBuzz**
- **MenuPages**
- **NewYork**
- **NYMag**
- **NYTimes**
- **OpenTable**
- **SavoryCities**
- **TasteSpace**
- **TimeOut**
- **VillageVoice**

Data preprocessing

Una prima analisi mostra come le modalità di recupero dei dati da parte dell'autore abbiano generato dei dataset differenti per schema e per frammentazione verticale, anche sulle stesse fonti. Dunque, prima di applicare le tecniche di record linkage, i dati sono stati standardizzati, eseguendo operazioni di riallineamento dello schema, rimozione dei duplicati ed infine join dei dataset per fonte.
La fase di preprocessing, nella quale sono incluse la pulizia dei dati e la standardizzazione, sono importanti poiché potrebbero aumentare l'accuratezza del record linkage.
In dettaglio le fasi del preprocessing sono state:

1. **Separazione e ragruppamento dei dati per fonte**.
   Ciò ha mostrato le differenze sullo schema, utilizzate nelle fasi successive per migliorare il raggruppamento dei dati.

2. **Separazione dei dati per fonte e data**.
   Utilizzando regex ad hoc per ogni fonte i dati sono stati separati per fonte e data, con drop dei duplicati, standardizzazione dell'encoding del testo (da *windows-1252* a *UTF-8*) e formato csv.

3. **Merge dei dataset per fonte**.
   I vari dataset risultanti dalla fase 2 sono stati riuniti per fonte con un join applicato su colonne definite a priori.

4. **Pulizia finale**.
   I dataset dunque sono stati ripuliti utilizzando delle funzioni specifiche per ogni fonte, con il compito di riallineare lo schema, rimuovere duplicati e caratteri superflui dai valori (es spazi e tabulazioni in testa e in coda, caratteri speciali ecc.).

Risultati

Nel caso di MenuPages, per esempio, inizialmente i dati erano presenti in un txt in questo formato ed erano uniti ai dati di altri siti web:

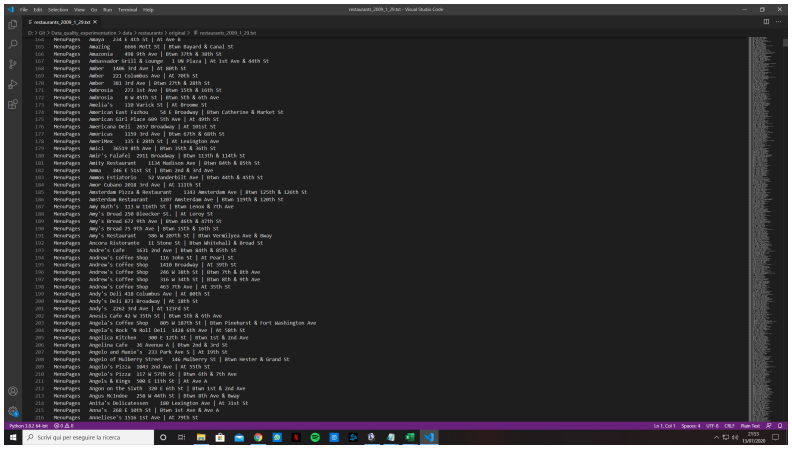

Fig. 1: *Dati di partenza di MenuPages*

a seguito del preprocessing invece otteniamo informazioni più chiare, aggiungendo l'informazione sull'indirizzo proveniente da Google come segue:

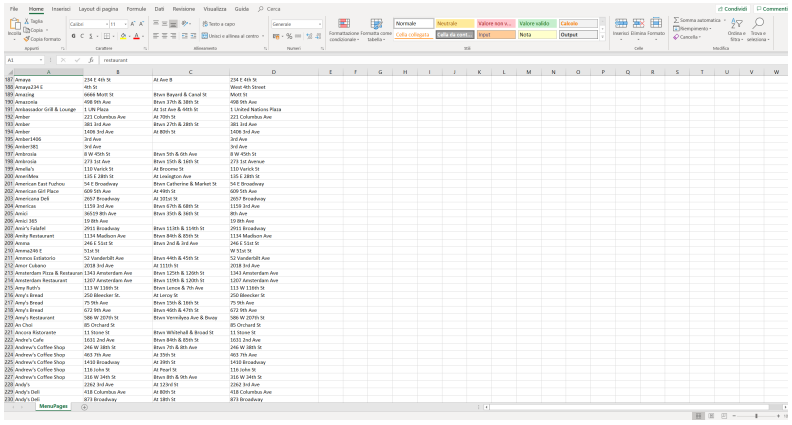

Fig. 2: *Dati dopo il preprocessing di MenuPages*

Record Linkage

Dalla fase di preprocessing abbiamo ottenuto in output 12 dataset, uno per ogni sito web presente nei file originali. Prima di procedere con l'attività di record linkage abbiamo deciso di introdurre una fase preliminare per costruire due dataset principali da poi linkare.

Fase preliminare: costruzione dataset

Analizzando i dataset ottenuti siamo riusciti ad individuare delle caratteristiche comuni in grado di distinguerli in due macro gruppi: Nei dataset *nomi dei dataset presenti in first df generator* si hanno esclusivamente il nome del ristorante e il quartiere (solo in *TimeOut* abbiamo in aggiunta l'indirizzo, che ci permetterà il linkage con l'altro dataset), nei dataset *nomi dei dataset presenti in second df generator* , invece, abbiamo nome del ristorante e indirizzo in tutti i file.

Per ogni insieme di dataset, dopo aver uniformato i nomi e l'ordine delle colonne presenti, abbiamo eseguito un append di tutti i dataframe generati importando i file csv. Successivamente sono stati rimossi i duplicati presenti basandoci solo sulle colonne comuni per tutti, ossia *restaurant* e *neighbourhood* nel primo caso e *restaurant* nel secondo (non è possibile utilizzare l'indirizzo in quanto non sono sempre scritti esattamente allo stesso modo nonostante rappresentino la stessa realtà)

Successivamente viene applicata la **deduplicazione**, ossia una tecnica usata per eliminare copie duplicate di dati ripetuti. È possibile applicarla attraverso la libreria *Record Linkage Toolkit* in quanto viene vista come un record linkage effettuato tra il database e se stesso. Come tecnica di *indexing* usiamo *sortedneighbourhood* (che illustreremo in dettaglio in seguito) ed effettuiamo la fare di *comparing* sugli attributi comuni nei rispettivi gruppi di dataset.

In questa fase iniziale per classificare i match ci siamo basati esclusivamente sul punteggio di score ottenuto dalla fase di comparing, in modo da poter includere con sicurezza record che differivano solo di poche lettere (ad esempio i ristoranti "Bubba Gump Shrimp Co." e "Bubba Gump Shrimp Company"), scegliendo come valore di soglia dello score

Una volta individuati i match sono stati unificati i dati in modo da creare un dataframe contenente le informazioni uniche utili per ogni match trovato. Per ottenere il dataset finale dal primo dataframe ottenuto come l'append dei csv importati sono state rimosse tutte le righe delle coppie presenti nei match e sono state aggiunti i record unificati contenenti le informazioni complete.

Da questa fase sono risultati due dataset, rispettivamente di 7524 e di 13379 record, invece dei record totali derivati dall'unione di tutti i 12 file iniziali.

Fase principale: record linkage

Per la fase di record linkage abbiamo seguito la seguente pipeline:

I passi principali seguiti sono i seguenti:

- **Indexing** - permette di creare coppie di record, denominate *candidate links*. Nel nostro algoritmo abbiamo implementato tre tecniche di indexing:
  - *full*, che crea le coppie effettuando il prodotto cartesiano dei due dataset, motivo per il quale è molto lungo e sconsigliato.
  - *blocking*, che permette di creare le coppie basandosi su una o più variabili uguali.
  - *sortedneighbourhood*, da usare nel momento in presenza di dataset con un grande numero di errori di spelling nei valori.
- **Comparing** - viene usato per confrontare le coppie di record create nella fase di indexing, sfruttando diversi metodi di similarità. Nel nostro caso abbiamo effettuato un confronto fra stringhe testando alcuni dei vari metodi disponibili ('jaro', 'jarowinkler', 'levenshtein', 'lcs') e impostando un valore di soglia (tutti i confronti approssimativi di stringhe più alti o uguali a questa soglia valgono 1) oppure usato un confronto esatto nel momento in cui trattavamo l'attributo *addressGoogle*, che sappiamo essere uniforme.
- **Classification** - dove le coppie vengono classificate in matches, non-matches e possible matches. In particolare abbiamo deciso di applicare due algoritmi di apprendimento non supervisionato, non essendo in possesso di training data. Gli algoritmi adottati sono:

- *ECM Classifier* o *Expectation/Conditional Maxisation classifier*, un algoritmo probabilistico dove viene trovata iterativamente la massima probabilità (locale) o la massima stima a posteriori (MAP) dei parametri nei modelli statistici. L'iterazione EM si alterna tra l'esecuzione di un passaggio di aspettativa (E), che crea una funzione per l'aspettativa della verosimiglianza, e un passaggio di massimizzazione (M).

  - *KMeans Classifier*, algoritmo che suddivide le coppie di record in match e non-match ed ogni vettore di confronto appartiene al cluster con la media più vicina. L'algoritmo è calibrato per due cluster: un cluster di corrispondenza e un cluster di non corrispondenza.

- Evaluation - permette di verificare la qualità del linkage in termini di accuratezza, recall e F-score. Purtroppo non ci è stato possibile eseguire questa fase in quanto sprovvisti di *veri positivi*.

Risultati:

Gold Standard

Il gold standard fornito è un elenco di 467 rimossi da alcuni siti web con l'informazione riguardante il suo stato al momento della verifica (Y = aperto, N = chiuso).