Sistema di Data Integration di sorgenti eterogenee

Alessio Marinucci & Riccardo Felici

https://github.com/alemari7/Hw5_IDD

Indice

1 Introduzione **2**

2 Task 1: Data Analysis e Definizione dello Schema Mediato **3**
2.1 Data Analysis 3
2.2 Definizione dello Schema Mediato 3
2.2.1 Obiettivi dello Schema Mediato 3
2.2.2 Strategie di Creazione dello Schema Mediato 3

3 Task 2: Popolamento dello Schema Mediato **4**

4 Task 3: Record Linkage **4**
4.1 Definizione della Ground Truth 4
4.2 Confronto Diretto tra Record 4
4.3 Pairwise Matching 5
4.4 Confronto Diretto tra Dataset 5
4.5 Pairwise Matching 5
4.6 Risultati e Valutazione 6

1 Introduzione

L'integrazione di sorgenti dati eterogenee rappresenta una sfida fondamentale nel campo della gestione e analisi dei dati aziendali. La crescente disponibilità di dati provenienti da fonti diverse, con strutture e semantiche variabili, richiede lo sviluppo di tecniche e strumenti avanzati per garantire una rappresentazione unificata e consistente delle informazioni. Questo documento descrive un approccio sistematico per affrontare il problema, applicato alle sorgenti dati relative alle aziende presenti nel repository *homework*.
L'obiettivo principale è definire uno schema mediato che permetta di armonizzare le informazioni provenienti da diverse sorgenti, affrontando le principali eterogeneità strutturali e semantiche. Tale schema mediato deve essere sufficientemente ricco da includere almeno 20 attributi, coprendo gli aspetti chiave delle entità aziendali, e flessibile per essere utilizzabile in contesti applicativi concreti.
Il processo di integrazione e analisi è articolato in tre fasi principali:

1. **Definizione di uno schema mediato**: Creazione di uno schema che includa almeno 20 attributi, allineando gli schemi delle sorgenti. Questo passaggio può essere affrontato attraverso una soluzione manuale o tramite un approccio personalizzato basato sull'intelligenza artificiale (come ChatGPT). Tali strumenti consentono di ridurre il tempo e lo sforzo richiesti per l'allineamento e la gestione delle eterogeneità.

2. **Popolamento dello schema mediato**: Unificazione e mappatura dei dati provenienti dalle diverse sorgenti in conformità con lo schema mediato. Questa fase è cruciale per garantire la qualità e la consistenza dei dati nel processo di integrazione.

3. **Calcolo del Record Linkage**: Identificazione delle entità corrispondenti tra le sorgenti utilizzando strategie avanzate. Questo include:

    - Creazione di una *ground-truth* contenente almeno 100 coppie in matching, includendo casi particolarmente complessi.
    - Applicazione di almeno due strategie di *blocking* per migliorare l'efficienza del processo.
    - Calcolo del *pairwise matching* tramite strumenti quali la libreria Python Record Linkage Toolkit e soluzioni basate su reti neurali come DeepMatcher, analizzando e confrontando i risultati ottenuti.

Questo lavoro si pone come obiettivo non solo quello di integrare e analizzare i dati presenti nel repository, ma anche di offrire un approccio replicabile e scalabile per affrontare problematiche analoghe in ambiti aziendali e accademici. Vengono inoltre analizzate le prestazioni e l'accuratezza delle diverse combinazioni di strategie, fornendo una valutazione approfondita dei vantaggi e dei limiti di ciascun metodo.

2 Task 1: Data Analysis e Definizione dello Schema Mediato

2.1 Data Analysis

La fase iniziale di analisi dei dati ha previsto una serie di attività mirate a estrarre, trasformare e preparare i dati per le successive elaborazioni. Di seguito sono descritte le principali operazioni svolte.

- Estrazione e decompressione dei file dal formato zip;
- Conversione dei file `json`, `jsonl` e `xls` in formato `csv` per garantire una maggiore compatibilità con gli strumenti di analisi dati.

2.2 Definizione dello Schema Mediato

La definizione dello schema mediato rappresenta il passo successivo nel processo di integrazione dei dati. Lo scopo è di fornire una rappresentazione unificata e standardizzata delle informazioni, garantendo coerenza e compatibilità tra le diverse sorgenti.

2.2.1 Obiettivi dello Schema Mediato

L'obiettivo principale di questa fase è costruire uno schema che:

- Contenga almeno 20 attributi significativi, rappresentativi delle entità aziendali presenti nelle sorgenti.
- Sia in grado di gestire e armonizzare le principali eterogeneità strutturali (*e.g.*, differenze nei formati dei dati) e semantiche (*e.g.*, variazioni nei nomi o nei significati degli attributi).
- Sia flessibile e scalabile, per poter essere utilizzato in contesti applicativi futuri.

2.2.2 Strategie di Creazione dello Schema Mediato

Utilizzo di Python. Python è stato impiegato per definire la lista degli attributi presenti nei vari dataset. Questo processo è stato supportato dall'analisi sintattica dei dati, mostrando la lista dei file in cui sono presenti attributi con nomi o significati simili. Il risultato è stato memorizzato nel file `mediated_schema.json`.

Supporto di ChatGPT. ChatGPT 4.0 è stato utilizzato per:

- Suggerire nomi coerenti e univoci per gli attributi.
- Associare ogni attributo ai file specifici in cui è presente, generando il file `mediated_schema.csv`.

Questo approccio ha permesso di migliorare la coerenza dello schema.

Raffinamento Manuale. Gli attributi sono stati completati e verificati manualmente, per garantire la massima qualità dello schema mediato.

3 Task 2: Popolamento dello Schema Mediato

Il secondo task si concentra sul popolamento dello schema mediato precedentemente definito. Questa fase prevede l'estrazione dei dati rilevanti dalle sorgenti, la loro trasformazione e la successiva integrazione nello schema unificato. L'obiettivo principale è garantire che tutte le informazioni provenienti dalle diverse fonti siano mappate correttamente agli attributi dello schema mediato, mantenendo coerenza e completezza. Di seguito, in Figura 1, parte dell'output prodotto dal nostro processo.

Unified Attribute	AmbitionBox.csv	DDD-cbinsight.com.csv	DDD-teamblind.com.csv	MalPatSaj-forbes.com.csv	MalPatSaj-wikipedia.org.csv	campaignindia.csv	companiesMarketCap_dataset.csv	company_social_urls.csv	disfold.com.csv	ft.com.csv
Name	Name	name	name	Name	Name	BRAND NAME	name	Company	name	name
City	headquarters	city	locations		Headquarters				headquarters	
Country		country		Country			country			country
Industry	Industry	industry	industry		Industry	CATEGORY	categories			industry
Founded	Foundation Year	founded	Founded		Founded					founded
Rank						RANK	rank			
investors		investors								
website			website						link	link
Market value				Market Value			market cap		market_cap	

Figure 1: Output dello Schema Mediato

4 Task 3: Record Linkage

Il *Record Linkage* è il processo di identificazione di record che rappresentano la stessa entità in uno o più dataset. Questo task è fondamentale per garantire l'integrità dei dati durante l'integrazione e per ottimizzare i processi di analisi. Di seguito descriviamo le fasi principali del processo, le tecniche adottate e i risultati ottenuti.

4.1 Definizione della Ground Truth

Per valutare le performance del sistema di *Record Linkage*, è stata definita una *ground truth* composta da almeno 100 coppie di record. Ogni coppia è stata manualmente classificata come valida o non valida sulla base della similarità tra i record.

4.2 Confronto Diretto tra Record

Un confronto diretto tra record è stato effettuato utilizzando due approcci principali di Blocking:

- **Blocking fonetico:** Utilizzando un identificatore fonetico basato sui nomi delle aziende, i record sono stati raggruppati in cluster. Questo approccio si basa su algoritmi fonetici come *MetaPhone*.
- **Blocking con Embeddings:** Generazione di rappresentazioni vettoriali (*word embeddings*) dei nomi delle aziende utilizzando il modello `MiniLM-L6`. I cluster sono stati creati applicando l'algoritmo `DBSCAN` sulle rappresentazioni vettoriali.

4.3 Pairwise Matching

Il *pairwise matching* è stato eseguito all'interno dei cluster generati. Sono state implementate due tecniche principali:

- Calcolo della similarità tra le coppie di record all'interno di ciascun cluster. Solo le coppie con similarità maggiore o uguale a 0.5 sono state considerate valide.
- Utilizzo del framework `DeepMatcher` basato su reti neurali per identificare automaticamente le coppie valide. Il modello è stato addestrato sul dataset generato dal *blocking*, diviso in *train*, *validation* e *test set* con proporzione 60/20/20.

4.4 Confronto Diretto tra Dataset

Un confronto diretto tra dataset è stato effettuato utilizzando due approcci principali di Blocking:

- **Attribute-based:** Le coppie di record sono state confrontate utilizzando una combinazione di attributi specifici. Ogni attributo è stato associato a un peso, determinato in base alla sua importanza relativa, e la similarità complessiva è stata calcolata come media ponderata delle similarità sugli attributi.
- **Similarity-based:** La similarità globale tra coppie di record è stata calcolata utilizzando metriche di similarità, come la distanza di Jaccard o il coefficiente di cosine, applicate direttamente sui valori concatenati degli attributi dei record.

4.5 Pairwise Matching

Il *pairwise matching* è una fase cruciale nel processo di *Record Linkage*, in cui si confrontano coppie di record per determinare la loro similarità. L'obiettivo principale è confrontare il *blocking* basato sulla similarità e quello basato sugli attributi.
L'obiettivo di questa analisi è valutare quale approccio di *blocking* fornisca risultati migliori in termini di precisione, richiamo e F-measure, rispetto a una *ground truth* predefinita.
Il processo prevede i seguenti passaggi:

- Caricamento dei dati da file CSV.
- Confronto delle coppie candidate utilizzando la libreria `RecordLinkage`.

- Calcolo delle metriche di performance (*precision*, *recall*, e *F-measure*) confrontandole con la *ground truth*.

Sono stati calcolati i seguenti indicatori di performance:

- **Precision:** Proporzione di coppie valide tra quelle identificate.
- **Recall:** Proporzione di coppie valide identificate rispetto a quelle presenti nella *ground truth*.
- **F1 Score:** Media armonica di precision e recall.

4.6 Risultati e Valutazione

1. Nell'approccio del confronto tra record, i risultati delle tecniche descritte sono stati confontati con la *ground truth*. Nel caso del modello basato su `RecordLinkage` i valori estratti sono i seguenti:

   - Accuratezza del 51% per il modello basato su blocking fonetico.
   - Accuratezza del 81% per il modello basato su blocking tramite embeddings.

   Nel caso del modello basato su `DeepMatcher` i valori estratti sono i seguenti:

   - Accuratezza del 55% per il modello basato su blocking fonetico.
   - Accuratezza del 71% per il modello basato su blocking tramite embeddings.

2. Nel caso di confronto diretto tra dataset, l'approccio basato su attributi ha dimostrato una maggiore efficienza nei dataset con schema ben definito, mentre l'approccio basato su similarità si è rivelato più robusto in presenza di dati eterogenei o rumorosi. Di seguito i risultati:

Approach	**Precision**	**Recall**	**F-Measure**
Similarity-based	1.0	0.44	0.61
Attribute-based	1.0	0.65	0.79

Table 1: Risultati di precision, recall e F-measure per approcci basati su similarità e attributi.