Data integration

Integrazione Dati e Costruzione di uno Schema Mediato

[Gabriele Rizzitiello]

Introduzione

Questa relazione documenta il processo di integrazione dati condotto come parte dell'homework 5, focalizzandosi sulla creazione di uno schema mediato. L'obiettivo principale era quello di unificare e rendere coerenti i dati provenienti da diverse fonti (sorgenti file .csv, .json, .jsonl, .xls), sfruttando sia tecniche manuali che l'ausilio di un Large Language Model (LLM). Il lavoro è stato suddiviso in fasi, partendo dall'estrazione degli attributi fino alla validazione finale, portando alla realizzazione di un dataset univoco basato sullo schema mediato ottenuto.
Infine sono stati considerati metodi di blocking distinti in combinazione con tecniche di pair-wise matching che confrontate con una ground-truth generata da noi hanno portato a diversi risultati che verranno analizzati nelle sezioni finali di questa relazione.

1. Realizzazione dello Schema Mediato: Approccio Ibrido

La costruzione dello schema mediato è stata effettuata con un approccio ibrido che combina l'accuratezza di un'analisi manuale con la scalabilità offerta dai LLM, nello specifico ChatGPT. Questa metodologia ha permesso di mitigare le difficoltà tipiche dell'integrazione dati.

- **Approccio Manuale:**
  - *Pro:* Elevata precisione semantica e controllo completo sul processo.
  - *Contro:* Richiede tempo e risorse umane considerevoli, limitandone la scalabilità.
- **Approccio LLM:**
  - *Pro:* Alta scalabilità, rapida elaborazione e riduzione del carico di lavoro manuale.
  - *Contro:* Possibili errori di interpretazione semantica e difficoltà nella gestione delle ambiguità.

L'integrazione di questi due approcci ha consentito di sfruttare i vantaggi di ciascuno, raggiungendo un buon compromesso tra efficacia ed efficienza.

```python
import os
import importlib
import pandas as pd

# Lista delle colonne dello schema mediato
mediated_schema_columns = [
    "Company ID", "Company Name", "Rank/Merit", "2010 Rank",
    "Annual Revenue", "Net Income", "Annual Results Year End",
    "Total Assets", "Total Liabilities", "Net Equity", "Headquarters Address",
    "Headquarters City", "Headquarters Country", "Headquarters Sub Region",
    "Headquarters Continent", "Headquarters Region", "Industry",
    "Business Sector(s)", "SIC Code", "EMTAK Code", "NACE Code",
    "Legal Form", "Foundation Date", "Join Date",
    "Company Number", "HHID", "CEO", "Founders", "Investors",
    "Official Website", "Market Valuation","Share Price", "Change 1 Day",
    "Change 1 Year", "Total Raised", "Company National ID",
    "Number of Employees", "Company Status", "Social Media - Facebook",
    "Social Media - Twitter", "Social Media - Instagram", "Social Media - Pinterest",
    "Ownership", "Main Market", "Notes", "Source", "Trade Name", "Postalcode"
]
```

2. Fasi di Integrazione Dati

- **Fase 1: Estrazione e Unificazione Iniziale**
  In questa fase, si è proceduto all'estrazione delle intestazioni delle tabelle o delle sorgenti dati, identificando i nomi degli attributi. Questo passaggio ha comportato la lettura di file di testo o il parsing dei file specifici. Successivamente, un LLM ha proposto un primo schema mediato, contenente un insieme di attributi unificati. Il prompt fornito a ChatGPT è stato mirato per rendere ibrido l'approccio e per ridurre la complessità semantica, individuando potenziali corrispondenze tra attributi.

- **Fase 2: Controllo e Correzione**
  Lo schema mediato generato dal LLM è stato sottoposto a un'analisi supervisionata. Durante questa fase, sono state identificate mappature errate, come l'unificazione di colonne non semanticamente equivalenti. In seguito, è stato fornito un secondo prompt al LLM, con l'obiettivo di affinare la logica di unificazione e garantire la correttezza semantica del modello.

- **Fase 3: Ricerca di una Mappatura Esaustiva**
  Si è verificato che alcuni attributi delle tabelle originali non erano stati inclusi nello schema mediato iniziale. Di conseguenza, lo schema mediato è stato espanso con nuove colonne per garantire una copertura completa di tutti gli attributi. A tale scopo, sono stati creati prompt specifici indirizzati al LLM, tra cui:

  “Prompt 1”
  "Unifica i seguenti attributi, riguardanti aziende, in uno schema mediato. Combina gli attributi semanticamente simili assegnando nomi coerenti, evidenziando possibili ambiguità"

  “Prompt 2”
  "È necessario unificare le colonne solo quando il loro contenuto informativo è lo stesso. Ad esempio, unire i social in un'unica colonna non è corretto, in realtà ogni social potrebbe avere il suo link diverso."

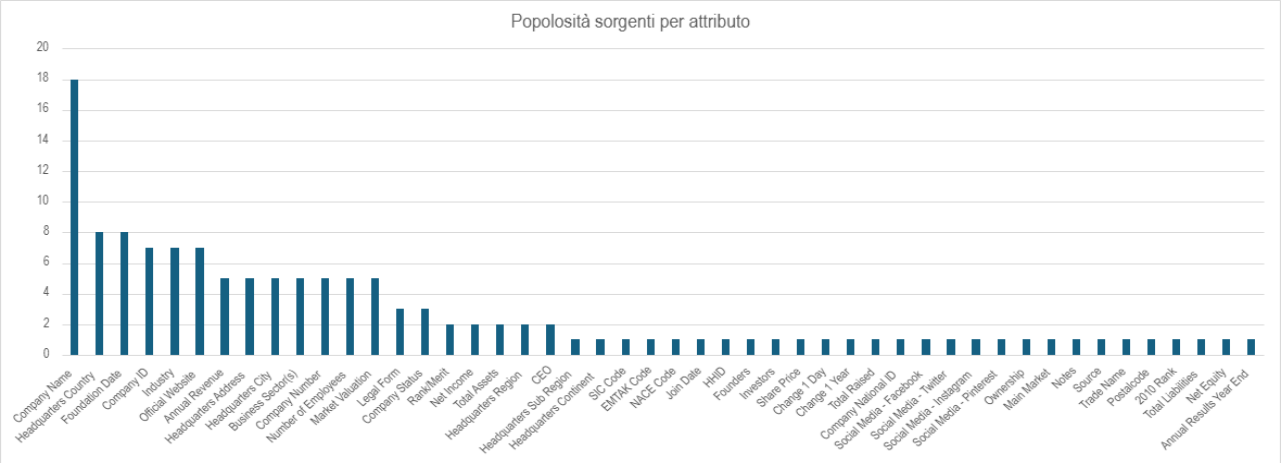
  In aggiunta, l’ LLM è stato istruito in modo da mappare tutti gli attributi delle tabelle di input con il modello, aggiungendo una nuova colonna nel caso di mancanza di corrispondenza.

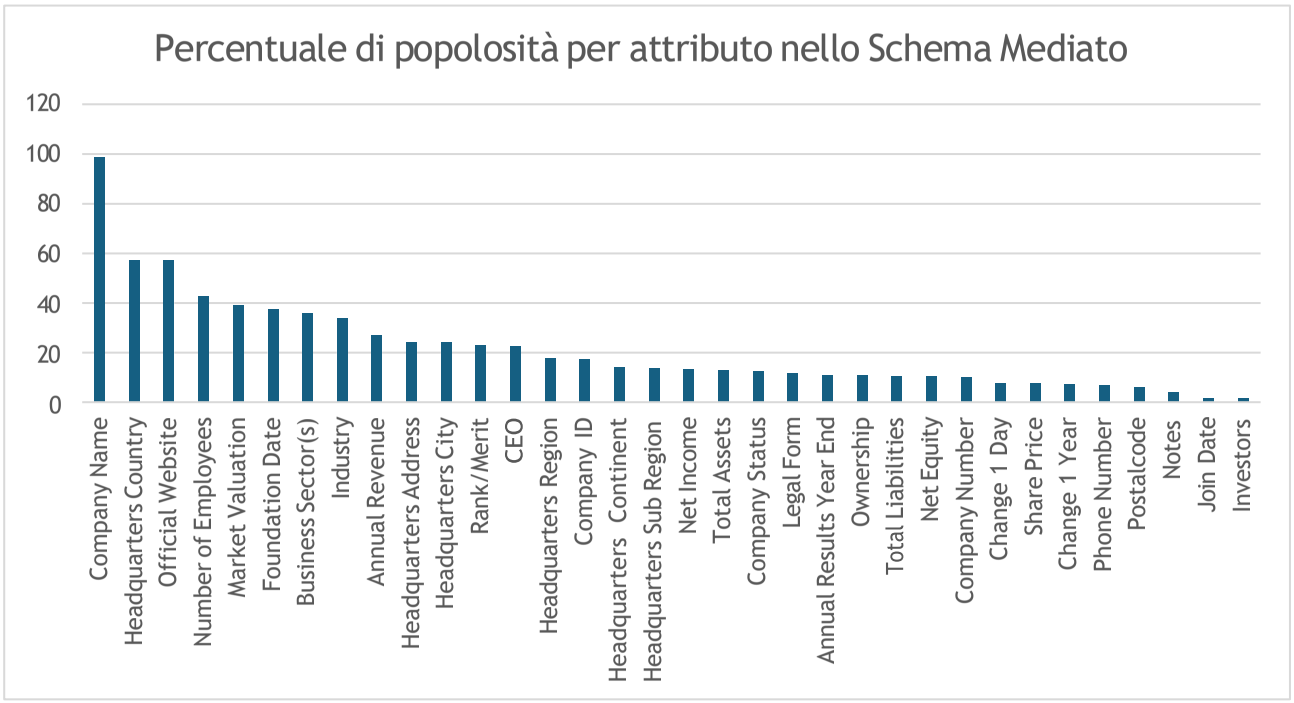
- **Fase 4: Validazione, Pulizia e Ottimizzazione**
  L'ultimo passaggio ha previsto una validazione manuale dello schema, al fine di identificare e rimuovere eventuali duplicati. Inoltre, per migliorare la precisione, sono state corrette alcune mappature e unificati i duplicati sotto un unico attributo. La validità dello schema è stata confermata tramite il comando df.info di Pandas, che ha mostrato le colonne effettivamente vuote. Al termine di questa fase, sono stati considerati **48** attributi totali.

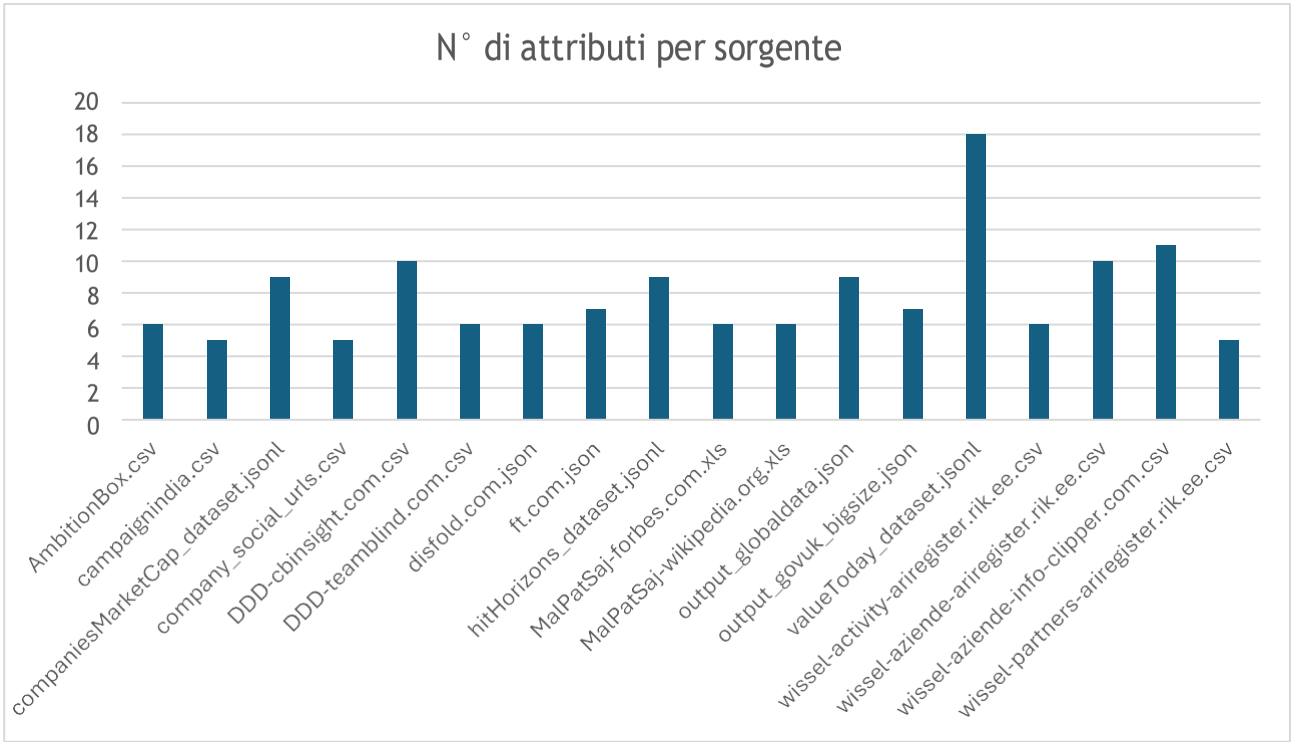
3. Popolamento dello Schema Mediato

Una volta validato e ottimizzato, lo schema mediato è stato popolato con i dati provenienti dalle sorgenti disponibili. Questo passaggio ha consentito di creare una rappresentazione coerente e centralizzata di tutte le informazioni, mantenendo l'integrità dei dati originali e garantendo la corretta relazione tra le informazioni.

Headquarters Country	Headquarters Sub Region	Headquarters Continent	Headquarters Region	Industry
India			Maharashtra	IT Services & Consulting
Ireland				IT Services & Consulting
United States (USA)				IT Services & Consulting
India			Maharashtra	Banking
India			Maharashtra	Banking
India			Karnataka	IT Services & Consulting
India			Karnataka	IT Services & Consulting
France				IT Services & Consulting
India			Maharashtra	IT Services & Consulting
United States (USA)			New York	IT Services & Consulting
				IT Services & Consulting
United States			Washington	Internet
India			Maharashtra	Banking
United States			New York	Software Product
United States (USA)			California	BPO
India			Maharashtra	Telecom
India			Maharashtra	Engineering & Construction
India			Gujrat	NBFC




Percentuale di popolosità per attributo nello Schema Mediato
120
100
80
60
40
20
0
Company Name
Headquarters Country
Official Website
Number of Employees
Market Valuation
Foundation Date
Business Sector(s)
Industry
Annual Revenue
Headquarters Address
Headquarters City
Rank/Merit
CEO
Headquarters Region
Company ID
Headquarters Continent
Headquarters Sub Region
Net Income
Total Assets
Company Status
Legal Form
Annual Results Year End
Ownership
Total Liabilities
Net Equity
Company Number
Change 1 Day
Share Price
Change 1 Year
Phone Number
Postalcode
Notes
Join Date
Investors


N° di attributi per sorgente
20
18
16
14
12
10
8
6
4
2
0
AmbitionBox.csv
campaignindia.csv
companiesMarketCap_dataset.jsonl
company_social_urls.csv
DDD-cbinsight.com.csv
DDD-teamblind.com.csv
disfold.com.json
ft.com.json
hitHorizons_dataset.jsonl
MalPatSaj-forbes.com.xls
MalPatSaj-wikipedia.org.xls
output_globaldata.json
output_govuk_bigsize.json
valueToday_dataset.jsonl
wissel-activity-ariregister.rik.ee.csv
wissel-aziende-ariregister.rik.ee.csv
wissel-aziende-info-clipper.com.csv
wissel-partners-ariregister.rik.ee.csv

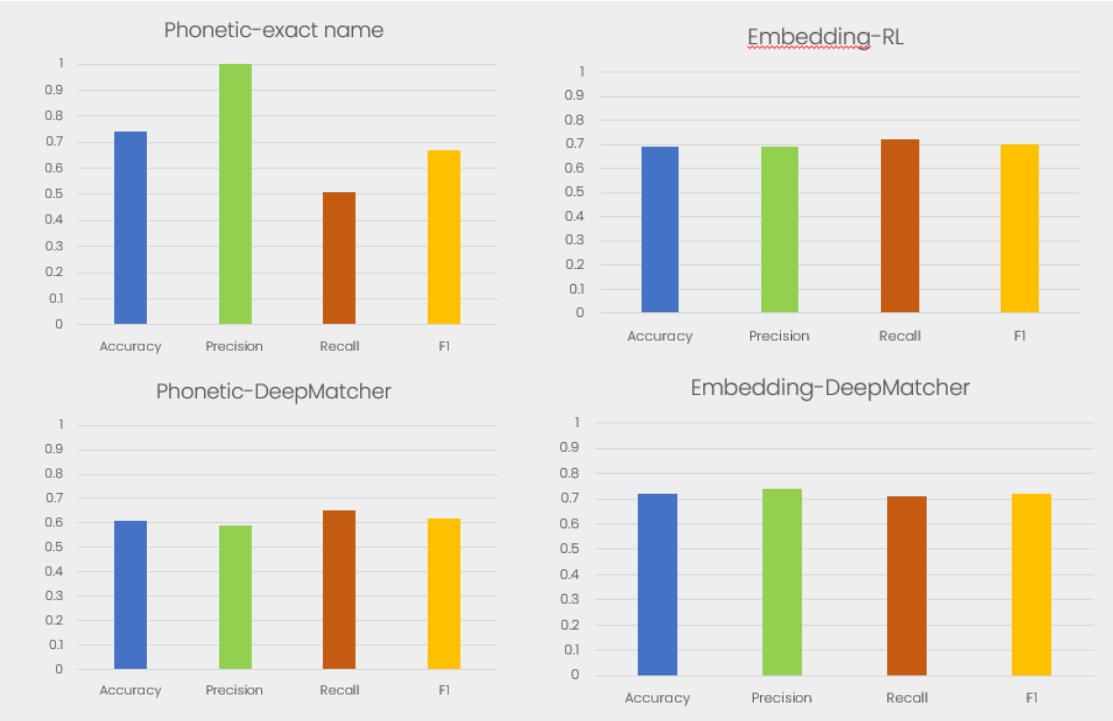
4. Pipeline del Sistema

Il processo di integrazione dati è stato organizzato in una pipeline, suddivisa in diverse fasi:

- **Dataset sullo schema mediato**: I dati vengono integrati e uniti in base allo schema mediato ottenuto precedentemente.
- **Fase di blocking**: In questa fase vengono applicati metodi di blocking tramite l'applicazione di un codice fonetico (es. Soundex) sul campo "Company Name" e con la creazione di vettori di embedding sui campi "Company Name" e "Country".
- **Fase di pair-wise matching**: Applicazione di tecniche per calcolare la similarità tra record, utilizzando il blocking.
- **Clusterizzazione**: I record sono raggruppati sulla base della similarità.
- **Confronto con la ground-truth**: I risultati ottenuti vengono confrontati con una ground-truth definita, al fine di valutare le performance del modello.

5. Valutazione delle Performance

Per la valutazione del sistema, sono state utilizzate diverse metriche quali Accuracy, Precision, Recall e F1-Score, applicate alle tecniche di matching fonetico e su embedding. Per la validazione dei risultati si è proceduto con dei test sia su DeepMatcher che su RL. I risultati mostrano buone performance per tutti i modelli, ma in particolare si nota un miglioramento dell'utilizzo di DeepMatcher insieme agli embedding.
Per l'embedding e phonetic abbiamo eseguito una visualizzazione delle relazioni per evidenziare la similarità semantica delle entità, dove possiamo osservare come aziende che soddisfano una soglia di similarità sui campi "Company Name" e "C. Name e Country", si ritrovano nello stesso cluster.



6. Conclusioni e Sviluppi Futuri

Il processo di integrazione dati descritto ha dimostrato l'efficacia di un approccio ibrido, che sfrutta sia l'accuratezza dell'analisi manuale che la scalabilità dei LLM. Nonostante i risultati promettenti, sono emerse alcune limitazioni.

- **Difficoltà:** Le principali difficoltà riscontrate sono state di tipo computazionale, in particolare per quanto riguarda la creazione di embedding di alta qualità.
- **Conclusioni:** L'uso combinato di embedding e DeepMatcher ha mostrato capacità prestazionali superiori rispetto ai soli metodi fonetici.
- **Sfide future:** L'impiego di embedding su un numero maggiore di campi e l'ottimizzazione dell'architettura della pipeline possono portare a miglioramenti significativi.

Conclusioni Finali

Questo lavoro ha dimostrato la fattibilità di un processo di integrazione dati che, partendo da una fase di estrazione iniziale, ha portato alla creazione di un solido schema mediato. L'approccio ibrido si è rivelato vincente per la sua capacità di unire accuratezza e scalabilità. I risultati ottenuti sono un punto di partenza per ulteriori ricerche e sviluppi, con l'obiettivo di ottenere un'integrazione dati sempre più efficiente e accurata.