Proposte di Tesi Magistrale e Progetto Individuale di GA AA 2022-2023

Questo documento riporta alcune proposte di argomenti di tesi magistrale e argomenti di progetto individuale per il corso di Graph analytics

NOTA: Per tesine riguardanti il corso Sistemi Informativi e Semantic Web si rimanda a https://docs.google.com/document/d/1ScQXFO78SuxZksDWYW0aCtexGTTxAWESTDVxzYFj1Zg/edit?usp=sharing

LEGENDA:
[GA] = argomento per tesina del corso Graph Analytics
[M]= argomento per tirocinio di laurea magistrale

Progetti su Knowledge Graph

Creazione di un UNIMORE Knowledge Graph [GA] [M]

DATI A DISPOSIZIONE:
Dati docenti UNIMORE: pagina http://personale.unimore.it/ , https://iris.unimore.it/
Settori: https://it.wikipedia.org/wiki/Settori_ERC

SCOPO: Lo scopo è quello di studiare e costruire un Knowledge Graph per il personale ricercatore di UNIMORE che contenga le informazioni del personale, principalmente docente: nome, cognome, afferenza, SSD, pubblicazioni, topic, co-work…

Su tale knowledge graph saranno effettuate poi diverse operazioni di analisi, quali a titolo esemplificativo:

- cercare gruppi di persone che collaborano nella pubblicazione articoli (banale - gruppi di ricerca)
- cercare topic tra diversi gruppi - estrarre i topic /keyword dalle pubblicazioni e valutare gruppi di pubblicazioni che fanno riferimento a topic similari

Articoli di riferimento:

- Challenges of Linking Organizational Information in Open Government Data to Knowledge Graphs https://arxiv.org/abs/2008.06232
- Challenges and Innovations in Building a Product Knowledge Graph https://dl.acm.org/doi/10.1145/3219819.3219938
- Challenges of Linking Organizational Information in Open Government Data to Knowledge Graphs https://www.researchgate.net/publication/346416824_Challenges_of_Linking_Organizational_Information_in_Open_Government_Data_to_Knowledge_Graphs
- Creating a Scholarly Knowledge Graph from Survey Article Tables https://arxiv.org/pdf/2012.00456.pdf
- A User Interface for Exploring and Querying Knowledge Graphs (Extended Abstract) https://www.ijcai.org/Proceedings/2020/0666.pdf
- Behnam Rahdari, Peter Brusilovsky, *Building a Knowledge Graph for Recommending Experts*. 1st International Workshop on Challenges and Experiences from Data Integration to Knowledge Graphs co-located with the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019), Anchorage, Alaska, August 5, 2019. CEUR Workshop Proceedings 2512. http://ceur-ws.org/Vol-2512/paper3.pdf

GraphDB Integration for Enhanced Social Network Analysis [GA] [M]

DATI A DISPOSIZIONE: Vedi appendice A [dati da varie piattaforme e reti sociali, compresi profili utente, relazioni di amicizia, interessi e attività]

OBIETTIVO: Sviluppare un sistema di integrazione avanzato per i Graph Database, che permetta di analizzare le reti sociali in modo più efficiente. Il progetto mira a gestire grandi quantità di dati eterogenei e a identificare modelli, comunità e influenzatori all'interno delle reti sociali.

Risultati attesi: Implementazione di un framework scalabile per l'integrazione dei dati sociali nei Graph Database. Il sistema consentirà analisi complesse, come la rilevazione di comunità, l'identificazione di pattern di influenza e l'individuazione di tendenze all'interno delle reti sociali.

Esplorazione e studio di Knowledge Graph dinamici [GA] [M]

INTRO:I KG sono basi di conoscenza che descrivono entità/concetti (ad esempio, luoghi, persone, artefatti) e li rappresentano utilizzando la struttura flessibile di un grafo.
I fatti sono spesso estratti da conoscenze enciclopediche, come Wikipedia, o archivi strutturati esistenti (ad esempio Wikidata), o anche da fonti non strutturate come post sui social media (ad esempio Facebook Graph). Ad esempio, è probabile che un KG contenente informazioni sulle aziende includa concetti sulle società, i loro fondatori, persone con ruoli chiave, sedi centrali o decentrate, numero di dipendenti, ecc. Tuttavia, i fatti relativi a entità o concetti che stanno cambiando dinamicamente nel tempo di solito mancano o diventano presto obsoleti.
Le dinamiche in evoluzione degli eventi del mondo reale di solito non si riflettono nelle basi di conoscenza. Pertanto, i repository attuali tendono a rappresentare solo istantanee statiche di entità del mondo reale, ignorando i loro cambiamenti nel tempo.

DATI A DISPOSIZIONE: **Vedi appendice A**

OBIETTIVO: questo progetto mira ad esplorare soluzioni per la gestione degli **aspetti temporali** e in **evoluzione** dei KG e a sfruttare tali caratteristiche per un'analisi più approfondita dei dati.

Man T., Vodyaho A., Ignatov D.I., Kulikov I., Zhukova N., Synthesis of multilevel knowledge graphs: Methods and technologies for dynamic networks, (2023) Engineering Applications of Artificial Intelligence, 123, art. no. 106244

DOI: 10.1016/j.engappai.2023.106244

Progetti sull'analisi del testo, tecniche NLP

Text classification: comparing feature-vector approach with graph based approach [GA] → assegnata a Elisa Tomisani 16/06/2023 - terminata

DATI A DISPOSIZIONE:

- Dataset DICE (Dataset of Italian Crime Event news) https://github.com/federicarollo/Italian-Crime-News

SCOPO: Confrontare tecniche di text classification che utilizzano i feature vector con quelle che utilizzano le rappresentazioni di dati testuali basate sui grafi.

MATERIALE A DISPOSIZIONE:

- Text Analysis Using Different Graph-Based Representations: https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2551/2372
- Text Classification using Graph Mining-based Feature Extraction: https://intranet.csc.liv.ac.uk/~frans/PostScriptFiles/ai09jiang.pdf

UrbanGraph: grafo di città, comuni, vie e attività commerciali [GA]

INTRO: quando si analizzano eventi che hanno un riferimento nello spazio, avere una gerarchia dei luoghi può aiutare ad analizzare meglio i dati a disposizione.
Per esempio, pensiamo di avere i riferimenti ai luoghi in cui sono avvenuti furti di bici nell'ultimo mese. Se uno di questi furti è avvenuto al supermercato "Gelsi" e un altro furto è avvenuto al Dipartimento DIEF, attraverso la gerarchia possiamo inferire che due furti sono avvenuti nel quartiere "Punta" perchè entrambi i luoghi si trovano in quel quartiere.
Una valida fonte dati è OpenStreetMap (OSM).

SCOPO: costruire la gerarchia di comuni, frazioni e quartieri della provincia di Modena, per

ognuno di essi inserire nella gerarchia le sue vie e le principali attività presenti (amenity in OSM).

DATI A DISPOSIZIONE / MATERIALE:
- OpenStreetMap https://www.openstreetmap.org/
- tool di interrogazione di OSM https://overpass-turbo.eu/
- https://wiki.openstreetmap.org/wiki/OSMonto

CrimeGraph: Creazione di un KG di crimini su cui svolgere analisi avanzate [GA] [M] → assegnata a Vincenzo Macellaro 19/06/2023 - terminata

INTRO: il dataset DICE è stato parzialmente annotato manualmente per mettere in evidenza gli oggetti rubati negli eventi di tipo furto, le vittime e gli autori, e i luoghi intesi sia come nomi di città, comuni, frazioni e quartieri, sia come nomi di vie, piazze e attività commerciali.

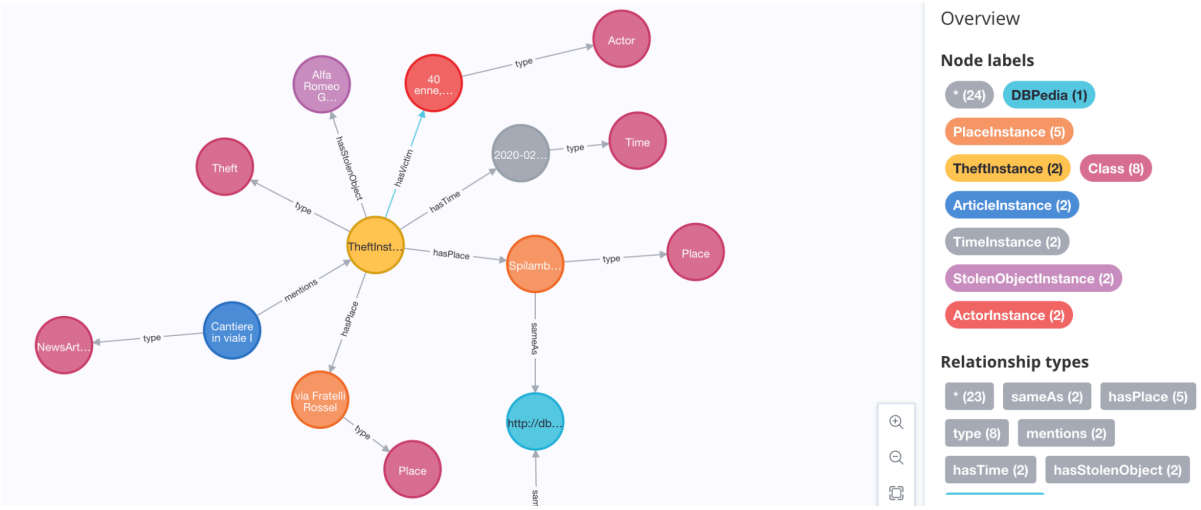
SCOPO: modellazione sotto forma di grafo dei dati annotati. Le analisi da effettuare verranno definite insieme allo/a studente/ssa che sceglierà di svolgere questa tesina.

DATI A DISPOSIZIONE:
- Dataset DICE (Dataset of Italian Crime Event news)
https://github.com/federicarollo/Italian-Crime-News

MATERIALE:
- Rollo, F., Po, L. (2022). Knowledge Graphs for Community Detection in Textual Data. In: Villazón-Terrazas, B., Ortiz-Rodriguez, F., Tiwari, S., Sicilia, MA., Martín-Moncunill, D. (eds) Knowledge Graphs and Semantic Web. KGSWC 2022. Communications in Computer and Information Science, vol 1686. Springer, Cham.
https://doi.org/10.1007/978-3-031-21422-6_15
- Federica Rollo, Laura Po, (2023) Modeling Event-Centric Knowledge Graph for Crime Analysis on Online News, Semantic AI in Knowledge Graphs, Taylor \& Francis, CRC Press (disponibile su richiesta)
- ONTOLOGIA - CEM

★★CrimeGraph2: Raffinamento del KG di crimini, disambiguazione, scoperta di nuove relazioni [GA] [M]

INTRO: Tale tesi estende la precedente da poco completata che ha dato luogo alla creazione di un knowledge graph



Il KG modella il dataset DICE, in particolare i dati annotati su tale dataset: mettendo in evidenza gli oggetti rubati negli eventi di tipo furto, le vittime e gli autori, e i luoghi intesi sia come nomi di città, comuni, frazioni e quartieri, sia come nomi di vie, piazze e attività commerciali.

SCOPO: Il KG non è completo: mancano ancora alcune categorie di dati annotati. Inoltre si vogliono inserire i riferimenti a sorgenti esterne, quali ad es. Babelnet, contenenti termini che disambiguano gli oggetti e luoghi dei furti. Ad esempio se in un furto sono stati rubati 3 orologi (informazione già presente nel grafo), tale oggetto rubato sarà associato al termine “orologio” in Babelnet.
Una volta esteso quindi il KG, verranno svolte delle analisi per identificare correlazioni tra oggetti rubati, luoghi dei furti, etc.

DATI A DISPOSIZIONE:
- Dataset DICE (Dataset of Italian Crime Event news)
https://github.com/federicarollo/Italian-Crime-News
- Dump di neo4j contenente il KG

MATERIALE:
- Rollo, F., Po, L. (2022). Knowledge Graphs for Community Detection in Textual Data. In: Villazón-Terrazas, B., Ortiz-Rodriguez, F., Tiwari, S., Sicilia, MA., Martín-Moncunill, D. (eds) Knowledge Graphs and Semantic Web. KGSWC 2022. Communications in Computer and Information Science, vol 1686. Springer, Cham.
https://doi.org/10.1007/978-3-031-21422-6_15
- Federica Rollo, Laura Po, (2023) Modeling Event-Centric Knowledge Graph for Crime Analysis on Online News, Semantic AI in Knowledge Graphs, Taylor \& Francis, CRC Press (disponibile su richiesta)
- ONTOLOGIA - CEM

Data augmentation per la categorizzazione di testi con BERT [GA] [M]

INTRO: Categorizzare un documento testuale significa individuare l'argomento descritto nel testo. Si può fare categorizzazione in diversi modi, un'opzione è utilizzare i word embedding per ottenere una rappresentazione numerica dei documenti.

SCOPO: Questo progetto intende utilizzare un modello BERT per assegnare la categoria di crimine ad alcuni articoli di giornale che descrivono crimini avvenuti nella città di Modena. Visto che test precedenti hanno dimostrato un problema con lo sbilanciamento del dataset (ci sono 13 categorie totali e circa il 70% dei documenti parla di furti), si chiede di trovare una strategia per la data augmentation in BERT.

RIFERIMENTI per la data augmentation:
- https://www.sbert.net/examples/training/data_augmentation/README.html
- AUG-BERT: An Efficient Data Augmentation Algorithm for Text Classification https://link.springer.com/chapter/10.1007/978-981-13-9409-6_266#:~:text=Aug%2DBERT%20is%20a%20data,CNN%20with%20dropout%20are%20adopted.

Valutazione della similarità tra concetti in BabelNet [GA] [M] → assegnata a GIOVANNI MALAGUTI 31/07/2023

INTRO: BabelNet è una rete semantica multilingua che integra la conoscenza proveniente da diverse sorgenti eterogenee in diverse lingue, come Wikitionary, Wikipedia, WordNet e molte altre. La struttura di BabelNet, analogamente a WordNet, raggruppa le parole in insiemi di sinonimi, chiamati Babel synset, i quali sono collegati da relazioni semantiche che legano i concetti rappresentati dai synset: iperonimia, iponimia, meronimia, olonimia, ecc.

BabelNet consente di mettere in evidenza le relazioni semantico lessicali tra i synset, grazie alle quali è possibile stabilire quanto è grande la correlazione semantica tra i concetti. BabelNet è dotato di API in diversi linguaggi (Java, Python, ecc) con le quali è possibile interrogare la rete e recuperare determinati synset. A differenza di WordNet, però, non esiste nessuna libreria che permette di misurare la **similarità semantica tra due Babel synset.**

SCOPO: creazione di funzioni e di una piccola libreria che permetta di calcolare la similarità tra i synset di BabelNet.

RIFERIMENTI:
- Navigli, Roberto & Ponzetto, Simone. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. 216-225.
- Navigli, Roberto & Bevilacqua, Michele & Conia, Simone & Montagnini, Dario & Cecconi, Francesco. (2021). Ten Years of BabelNet: A Survey. 4559-4567. 10.24963/ijcai.2021/620.
- Pedersen, Ted & Patwardhan, Siddharth & Michelizzi, Jason. (2004). WordNet::Similarity - Measuring the Relatedness of Concepts
- https://www.geeksforgeeks.org/nlp-wupalmer-wordnet-similarity/
Reference paper about BabelNet

R. Navigli and S. Ponzetto, BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network.
Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250
Winner of the Artificial Intelligence Journal 2017 Prominent Paper Award

★ Scientific Conference recommendation system [GA] [M]

INTRO: Gli articoli scientifici sono pubblicati principalmente in conferenze internazionali oppure riviste scientifiche. Per le riviste esistono alcuni tool che permettono di suggerire agli autori dell'articolo a quale rivista sottomettere il proprio articolo a partire dal titolo e le parole chiave (ad esempio https://journalfinder.elsevier.com/).
Per le conferenze non esiste un tool del genere.

SCOPO: Obiettivo di questo progetto è sviluppare un web service che suggerisca una lista di conferenze appropriate all'articolo che si intende sottomettere sulla base degli articoli pubblicati nelle edizioni passate di quelle conferenze. Questo può essere fatto attraverso una indicizzazione per parole chiave degli articoli già pubblicati.

STEP EXTRA: ogni conferenza si tiene una volta all'anno ed ha un periodo di sottomissione, a cui segue un periodo di valutazione degli articoli e quindi la revisione e la ri-sottomissione degli stessi, in caso di valutazione positiva. Le date di riferimento di questi periodi si trovano sulla pagina web della conferenza. Un autore può essere interessato o meno ad una conferenza anche in base a quale sia la data della prima sottomissione, quindi il web service, dove possibile, deve contenere anche queste informazioni che si possono ottenere con un web crawler.

MATERIALE:
- sito del rating delle conferenze (https://scie.lcc.uma.es:8443/ratingSearch.jsf)

Progetti in ambito Smart City

Graph data science and machine learning applied on transport networks [M]

La rete a grafo per il trasporto multi modale permette di ottimizzare servizi di trasporto e analizzare i punti critici nella città o nei quartieri.

I progetti che possono essere sviluppati in tale ambito sono:

- studio e analisi delle migliori pratiche, standard e servizi esistenti per la gestione del traffico intermodale e dei diversi graph model
- utilizzo di big data, digital twin e tecnologie HPC per archiviare dati ed elaborare statistiche e simulazioni su set di dati scalabili
- implementazione di una rete a grafo per la città di Ferrara
- sviluppo di uno strumento per la visualizzazione efficace dei dati di mobilità
- implementazione di mappe urbane di sicurezza per le città di Modena e Ferrara che classifichino le strade e i percorsi in base alle condizioni di sicurezza per i pedoni e i ciclisti con una prospettiva di genere
- identificazione delle aree/quartieri meno connessi nel graph model e valutazione di possibili scenari che ottimizzano i trasferimenti
- sviluppo di routing tool per individuare il percorso multimodale più breve e il percorso più sicuro per i ciclisti che attraversano la città.

Tali progetti sono collegati al progetto ECOSIST-ER (Ecosystem for Sustainable Transition in Emilia-Romagna), finanziato dall'uninione Europea all'interno del programma NextGenrationEU, ed in particolare allo Spoke 4 - Smart mobility, housing and energy solutions for a carbon-neutral society

★ Integrazione del file GTFS di trasporto pubblico nel grafo stradale di Modena [GA] [M] -->assegnato a Paolo Attardi 20/06/2023 - terminata

DATI A DISPOSIZIONE: File GTFS del trasporto pubblico (descrizione formato GTFS https://developers.google.com/transit/gtfs?hl=it).

SCOPO: generare un database a grafo che integri i dati di trasporto pubblico e creare uno script python in grado di calcolare il percorso più breve. Esistono già delle librerie di python che consentono di convertire direttamente i file GTFS in strutture a grafo ed interrogarle (https://pypi.org/project/peartree/). L'idea sarebbe di generare poi una istanza di neo4j facendo un po' di refactoring ed applicare algoritmi di routing per calcolare il percorso più breve.
In alternativa al routing si possono effettuare altre analisi della topologia del grafo ottenuto e ulteriori analisi come la valutazione della centralità delle fermate o community detection dei percorsi.

MATERIALE:
http://kuanbutts.com/2018/03/15/comparative-routes-mpl/
https://github.com/ChiaraBachechi/publicTransportLines
https://pypi.org/project/peartree/
https://neo4j.com/videos/discover-neo4j-auradb-free-with-michael-and-alexander-18/

★ Utilizzo delle librerie SCIPY.SPATIAL e GeoPandas per studiare la relazione spaziale tra ciclabili, percorsi pedonali e strade [GA]

INTRO: SCIPY.SPATIAL è una libreria Python che permette di eseguire diverse funzionalità spaziali sui dati. L'idea è di sfruttare le potenzialità di questa libreria insieme a GeoPandas (che include la libreria shapely) ed eventualmente anche di altre librerie di Python per generare uno script che studi la relazione spaziale tra elementi delle rete di trasporto urbana.

Nel seguito elenchiamo 3 possibili sottoprogetti (la tesina ne dovrà affrontare solo 1):

a) Valutazione CICLABILI

DATI A DISPOSIZIONE: Le geometrie delle ciclabili sono disponibili sul geoportale dove è possibile trovare anche la geometria delle strade. I dati sulle aree verdi sono disponibili su OSM.

SCOPO: Studiare la relazione spaziale tra ciclabili e strade per la circolazione dei veicoli, valutare la qualità dell'ambiente che circonda la ciclabile (presenza in prevalenza di aree verdi, separazione dal traffico). Attribuire ad ogni tratto di ciclabile un punteggio basato sulla sua vicinanza alle aree verdi ed un diverso valore basato invece sulla vicinanza al traffico.

b) Valutazione PERCORSI PEDONALI

DATI A DISPOSIZIONE: Le geometrie dei percorsi pedonali sono disponibili su OSM così come le informazioni sulle aree verdi.

SCOPO: Studiare la relazione spaziale tra percorsi pedonali, strade e aree verdi. Attribuire ad ogni tratto di percorso pedonale un punteggio basato sulla sua vicinanza alle aree verdi ed un diverso valore basato invece sulla vicinanza al traffico.

c) Valutazione CONTESTO STRADALE

DATI A DISPOSIZIONE: posizione edifici e punti di interesse da OSM, rete stradale dal geoportale o da OSM

SCOPO: valutare la percentuale di area che circonda una strada che è composta da:

- altre strade
- parchi e aree verdi
- attività commerciali
- attività di svago

- edifici abitativi (se possibile)

MATERIALE:
https://github.com/ChiaraBachechi/CyclePathSecurityLevels
https://docs.scipy.org/doc/scipy/reference/spatial.html
https://www.w3schools.com/python/scipy/scipy_spatial_data.php
https://www.w3schools.com/python/scipy/scipy_interpolation.php
https://scikit-geometry.github.io/scikit-geometry/skeleton.html

Analisi del grafo stradale della città di Modena [GA]

DATI A DISPOSIZIONE: grafo stradale della città di modena (istanza di neo4j). Il grafo già contiene i punti di interesse presenti su OSM.

INTRO: Il grafo della rete stradale di Modena è già stato impelmentato in una rappresentazione che unisce primal e dual graph in una sola istanza. Non è necessario utilizzare entrambe le rappresentazioni ma è possibile scegliere solo quella che meglio si adatta allo scopo dell'analisi.
SCOPO: la tesi richiede inizialmente di studiare e comprendere la struttura del grafo esistente (vi verrà fornito un paper ed alcune slide in inglese che la descrivono). Successivamente verranno effettuate alcune analisi della struttura del grafo considerando sia la struttura della rete stradale che la presenza dei punti di interesse.
Esempi di possibili analisi:

- qualità delle interconnessioni tra i punti di interesse
- random walk nel grafo e valutazione della probabilità di ogni nodo di essere contenuto nel percorso

MATERIALE:
https://github.com/ChiaraBachechi/roadRouting
https://neo4j.com/docs/graph-data-science/current/algorithms/random-walk/

Progetti su qualità dell'aria e ambiente

1. Air quality **augmentation** [GA] [M]

INTRO: Utilizzare tecniche di data augmentation (vedere ultimo link) su dati della qualità dell'aria rilevati da sensori low-cost.

DATI A DISPOSIZONE: Dati rilevati da sensoristica low-cost (PM) e dati di riferimento di stazioni di monitoraggio Arpa.

SCOPO: La concentrazione di alcuni inquinanti nell'aria è strettamente influenzata dal contesto ambientale, inclusi fattori come le condizioni meteorologiche. Utilizzando i dati rilevati dai sensori low-cost, intendiamo applicare tecniche di data augmentation (anche applicate ai grafi) per ottenere un dataset più ampio su cui allenare reti neurali. Ciò ci consentirà di affrontare diverse sfide, come l'individuazione di anomalie, l'interpolazione dei dati mancanti, la riduzione degli errori correlati all'igroscopia e altro ancora. L'obiettivo finale è migliorare la precisione e l'affidabilità delle analisi e delle previsioni legate alla qualità dell'aria.

RIFERIMENTI:
Codice in Python

https://www.datacamp.com/tutorial/complete-guide-data-augmentation
https://www.wiseair.vision/
https://github.com/zhao-tong/graph-data-augmentation-papers

2. Applicare tecniche di **clustering** ai dati di qualità dell'aria urbana [GA] [M]

INTRO: applicare tecniche di clustering alle serie temporali delle concentrazioni di inquinanti rilevate dai sensori low-cost.

DATI A DISPOSIZONE: Dati rilevati da sensoristica low-cost (PM) con coordinate associate e dati di riferimento di stazioni di monitoraggio Arpa.

SCOPO: L'obbiettivo è utilizzare le tecniche di clustering per identificare contesti simili tra i sensori low-cost della qualità dell'aria installati sul territorio. Questo ci permetterà di raggruppare i sensori che mostrano pattern di concentrazione di inquinanti simili nel corso del tempo, contribuendo così a una migliore comprensione della distribuzione dell'inquinamento atmosferico e delle sue fonti all'interno dell'area di monitoraggio.

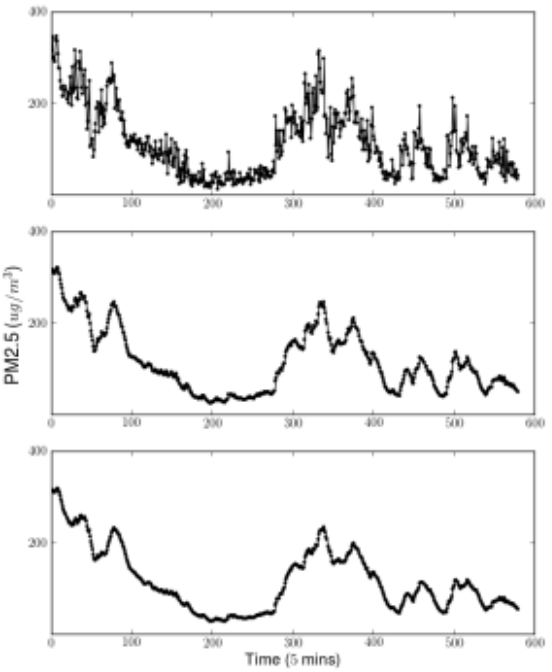
RIFERIMENTI:

Codice in Python

https://www.wiseair.vision/

Time Series Clustering — Deriving Trends and Archetypes from Sequential Data
Malaysia PM10 Air Quality Time Series Clustering Based on Dynamic Time Warping
A multi-variate time series clustering approach based on intermediate fusion: A case study in air pollution data imputation

3. Confronto di tecniche di **pre-processing** su dati di qualità dell'aria [GA] [M]



INTRO: Nell'ambito del monitoraggio della qualità dell'aria, i sensori low-cost forniscono un segnale grezzo che può essere affetto da fluttuazioni indesiderate causate dal rumore. Per migliorare la qualità dei dati, esistono diverse metodologie per correggere tali fluttuazioni. Questo progetto si propone di esplorare queste metodologie e sviluppare uno script che, a partire dai dati grezzi, generi una serie temporale "pulita".

DATI A DISPOSIZONE: Dati rilevati da sensoristica low-cost (PM e gas) e dati di riferimento di stazioni di monitoraggio Arpa.

SCOPO: L'obiettivo di questa attività progettuale è applicare e studiare diverse tecniche di pre-processing per ottenere dati di qualità superiore, escludendo le anomalie statistiche grezze. Ad esempio applicando soglie statistiche basate sulla deviazione standard, i quartili, ecc., oppure metodologie di smoothing (es. utilizzando finestre mobili basate sulla mediana). Tali tecniche saranno applicate, singoalarmente e consecutivamente, alle serie temporali relative ai gas e alle particelle atmosferiche (PM) provenienti dai dati dei sistemi Wiseair/Trafair. Lo scopo ultimo e studiare quali tecniche di pre-processing hanno un impatto migliorativo sui dati della qualità dell'aria in esame, che verranno poi utilizzati in successive analisi.

RIFERIMENTI:
http://sensys.acm.org/2014/papers/p251-cheng.pdf
https://www.wiseair.vision/
A multi-variate time series clustering approach based on intermediate fusion: A case study in air pollution data imputation

4. **Calibrazione** attraverso algoritmi di **transfer learning** [M]

INTRO: La tesi prevede di applicare modelli di calibrazione già sviluppati e analizzare poi i risultati ottenuti.

DATI A DISPOSIZONE: Dati rilevati da sensoristica low-cost (PM e gas) e dati di riferimento di stazioni di monitoraggio Arpa.

SCOPO: I sensori di qualità dell'aria per loro natura richiedono di essere calibrati singolarmente ed indipendentemente dagli altri. Con l'idea di sfruttare il modello creato per un sensore su un altro sensore applicando delle correzioni. Si pone anche l'obiettivo di valutare la differenza tra le osservazioni ottenute quando i sensori sono posti nello stesso posto, sia per quanto riguarda le osservazioni grezze che per quanto riguarda invece le osservazioni calibrate.

RIFERIMENTI:
CrossSense: Towards Cross-Site and Large-Scale WiFi Sensing
A Comprehensive Survey on Transfer Learning
An Improved Hybrid Transfer Learning-Based Deep Learning Model for PM2.5 Concentration Prediction
Data Cleaning to fine-tune a Transfer Learning approach for Air Quality Prediction

5. **Predizione** livelli dell'inquinamento futuro [GA] [M]

INTRO: La tesi prevede di applicare modelli di preidzione per prevedere il livello di inquinamento futuro (applicabile anche sui grafi).

DATI A DISPOSIZONE: Dati rilevati da sensoristica low-cost (PM e gas) e dati di riferimento di stazioni di monitoraggio Arpa.

SCOPO: Utilizzare tecniche di forecasting per prevedere il livello futuro dell'inquinamento nei punti indicati o nello spazio (interpolazione). I sensori sono associati alle coordinate spaziali, è possibile quindi utilizzare un grafo per la loro rappresentazione. Esistono tecniche di predizione applcabili nel caso dei grafi.

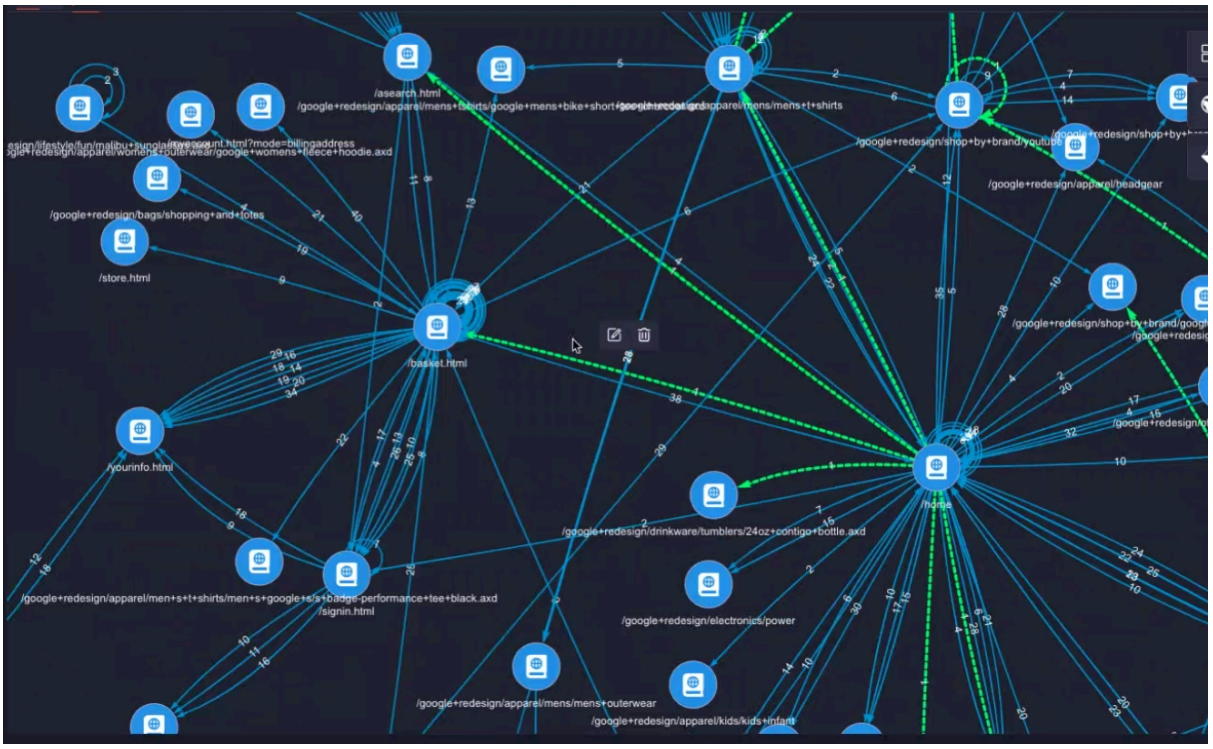
RIFERIMENTI:
Graph Neural Network for Air Quality Prediction: A Case Study in Madrid
A hybrid deep learning technology for PM2.5 air quality forecasting
Automatic outlier detection for time series: an application to sensor data

(*https://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=519616*)

Progetti in collaborazione con aziende

Individuare e implementare un nuovo algoritmo di Graph embedding - [M]

Grafo rete di navigazione(archi verdi il primo cambio pagina dell'utente)



Obiettivo cercare di classificare in maniera automatica la navigazione di un utente e raggrupparla. Tracciare l'anomalia nel comportamento utente per migliorare il sito web

Usare un nuovo algoritmo di graph embedding quindi trasformare il grafo in vettore e poi confrontarli per vedere la similarità

Trovare un nuovo algoritmo per fare graph embedding della porzione di grafo che interessa un utente

Quello che manca è un algoritmo che effettua l'embedding di una porzione di sottografo beh definito.

Il responsabile della graph analytics ha delle idee - vorrebbero studiare e progettare un algoritmo di graph embedding - la parte di implementazione potrebbero farla loro - studia le diverse possibilità

Miliardi di archi - rendere l'algoritmo performante.

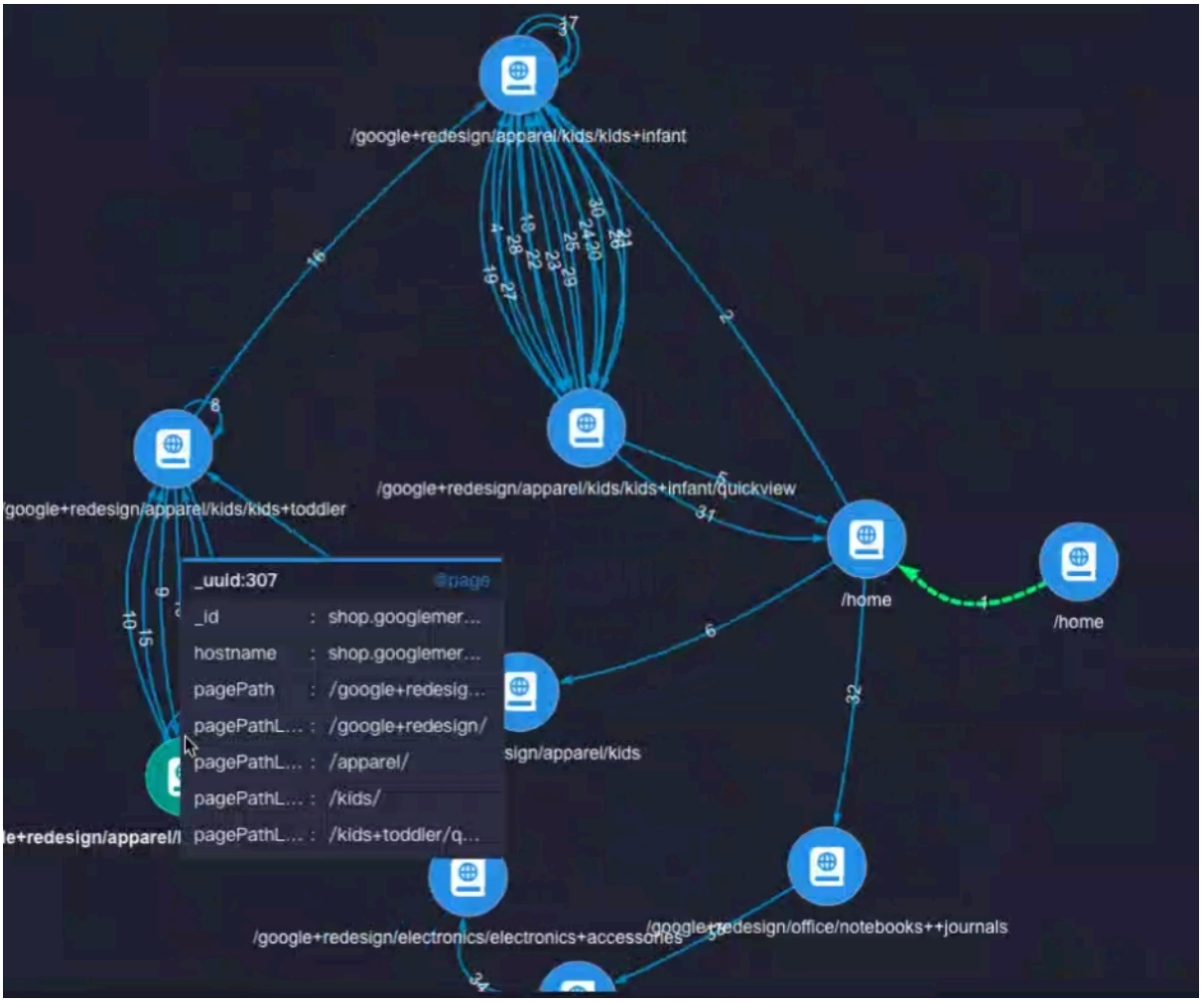


FIGURA: Attività di navigazione di un utente

Appendice A - Graph databases for testing purpose

- Stanford Large Network Dataset Collection (SNAP): SNAP fornisce una collezione di dataset di reti di grandi dimensioni, tra cui reti sociali, grafi web e reti di collaborazione. Puoi esplorare e scaricare vari dataset di grafi dal loro sito web: https://snap.stanford.edu/data/index.html
- Network Repository: Network Repository ospita una vasta gamma di dataset di reti provenienti da diversi ambiti, tra cui reti sociali, reti biologiche e reti di trasporto. Forniscono un'interfaccia user-friendly per cercare e scaricare dataset: http://networkrepository.com/
- Kaggle: Kaggle è una piattaforma per competizioni di data science che ospita anche un repository di dataset. Puoi cercare dataset relativi ai grafi su Kaggle e trovare dataset correlati a reti sociali, sistemi di raccomandazione e altre strutture di grafi: https://www.kaggle.com/datasets
- Open Graph Benchmark (OGB): OGB è una collezione di dataset di grafi e task di benchmark per l'apprendimento automatico su grafi. Si concentra su dataset di grafi del mondo reale sfidanti provenienti da diversi ambiti. Puoi trovare i dataset e le procedure di valutazione sul loro repository GitHub: https://github.com/snap-stanford/ogb
- UCI Network Data Repository: UCI Network Data Repository fornisce una collezione di dataset di reti, inclusi reti sociali, reti biologiche e reti di citazioni. Puoi accedere ai dataset e alle loro descrizioni dal loro sito web: http://networkdata.ics.uci.edu/index.php

Appendice B - Dataset in ambito mobilità (strade, ciclabili, flussi dati)

PORTALE Mobilità Regione Emilia Romagna
https://mobilita.regione.emilia-romagna.it/

PORTALE MINERVA - dataset
https://datacatalog.regione.emilia-romagna.it/catalogCTA/dataset

archivio strade
https://servizissiir.regione.emilia-romagna.it/ARS/

flussi traffico online (dati censiti dal Sistema regionale di rilevazione dei flussi di traffico dell'Emilia-Romagna)
https://servizissiir.regione.emilia-romagna.it/FlussiMTS/

piste ciclabili urbane (solo totale km)
https://datacatalog.regione.emilia-romagna.it/catalogCTA/dataset/piste-ciclabili-urbane-1532425158099-7227/resource/2e598292-da87-4e63-b58a-f9c3857d0be7

ciclovie regionali (WMS e pdf) - portale minerva -
https://datacatalog.regione.emilia-romagna.it/catalogCTA/dataset/ciclovie-regionali
https://mobilita.regione.emilia-romagna.it/allegati/prit/adottato/prit2025_carta_e-adottato.pdf/

PUMS
https://mobilita.regione.emilia-romagna.it/mobility-sostenibile/mobilita-sostenibile/pums-piani-urbani-per-la-mobilita-sostenibile

Open https://registry.opendata.aws/

Appendice C - Tool di visualizzazione

- https://flourish.studio/ - per creare grafici interattivi anche a partire da CSV
- https://rawgraphs.io/ - creato sfruttando la libreria D3.js, permette di creare grafici usando un'interfaccia

Riferimenti bibliografici:

Dirk Streeb, Mennatallah El-Assady, Daniel A. Keim, Min Chen, Why Visualize? Arguments for Visual Support in Decision Making, IEEE Computer Graphics and Applications, 2021 , 10.1109/MCG.2021.3055971 https://bit.ly/3cKVtxp

- Story telling
  https://www.gokantaloupe.com/blog/best-techniques-for-data-driven-storytelling
- Data Story telling Course
  https://www.pluralsight.com/courses/data-storytelling-moving-beyond-static-data-visualizations?aid=7010a000002BWqGAAW&promo=&utm_source=non_branded&utm_medium=digital_paid_search_google&utm_campaign=EMEA_Dynamic&utm_content=&gclid=Cj0KCQiAtJeNBhCVARIsANJUJ2Hvn8cPrOLYBa5U-d9bMbeJxnAcT-zrUty_2nG1ydVAw3lBVMpdoj4aAseNEALw_wcB