



UNIVERSITA' DEGLI STUDI DI PADOVA
Dipartimento di Ingegneria dell'Informazione

Corso di Laurea in Ingegneria Informatica

Relazione per la prova finale

Knowledge Graph Completion: Definizione, Metodi e Applicazioni

Relatrice: Prof.ssa Ornella Irrera

Laureando: Gianluca Compagno
Matricola n.: 2101760

ANNO ACCADEMICO 2025/2026



INTRODUZIONE

- I Knowledge Graph (KG) permettono di organizzare in modo efficiente informazioni e concetti appartenenti al mondo reale
- Un problema comune a quasi tutti i KG è quello dell'incompletezza, essa deriva dal fatto che esistono entità e relazioni che non sono state scoperte
- La scarsità di entità e relazioni causa una mancanza di informazioni rilevanti e limita l'efficacia del KG
- La soluzione è il Knowledge Graph Completion (KGC), esso permette di migliorare la qualità del KG completandolo con nuovi fatti non ancora osservati
- La necessità di KG affidabili per applicazioni finali focalizza l'attenzione su tecniche di KGC capaci di scoprire automaticamente nuovi fatti

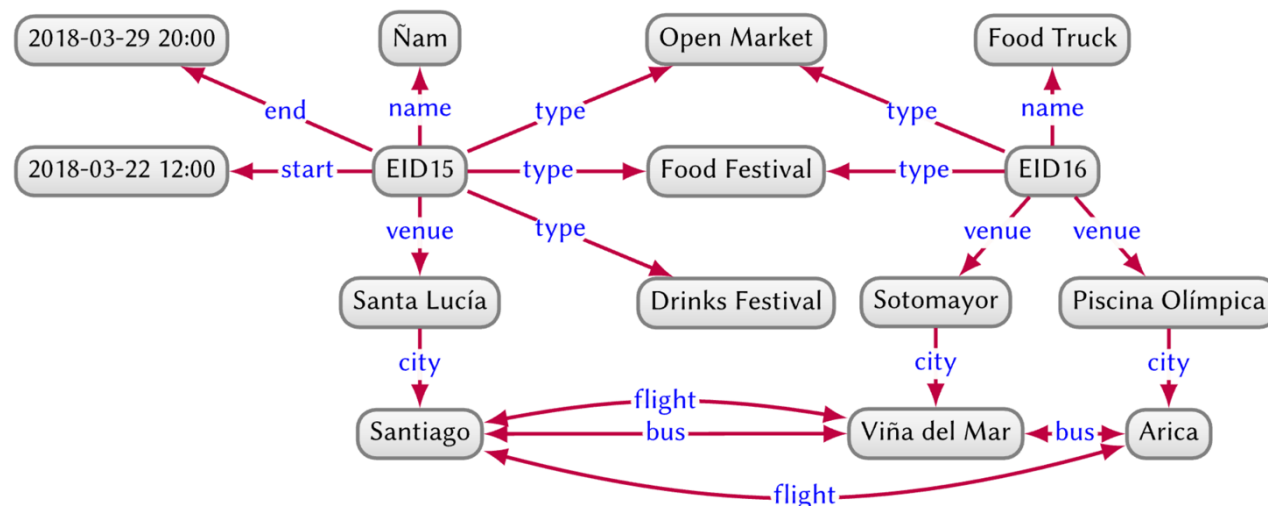


KNOWLEDGE GRAPH

- I KG sono grafi formati da nodi che rappresentano le entità e archi che rappresentano le relazioni tra entità
- OBIETTIVO: organizzare informazioni in modo efficiente
- STRUTTURA: un fatto corrisponde ad una tripla del tipo (*head, relation, tail*)
- PROBLEMA: incompletezza



- SOLUZIONE: Knowledge Graph Completion



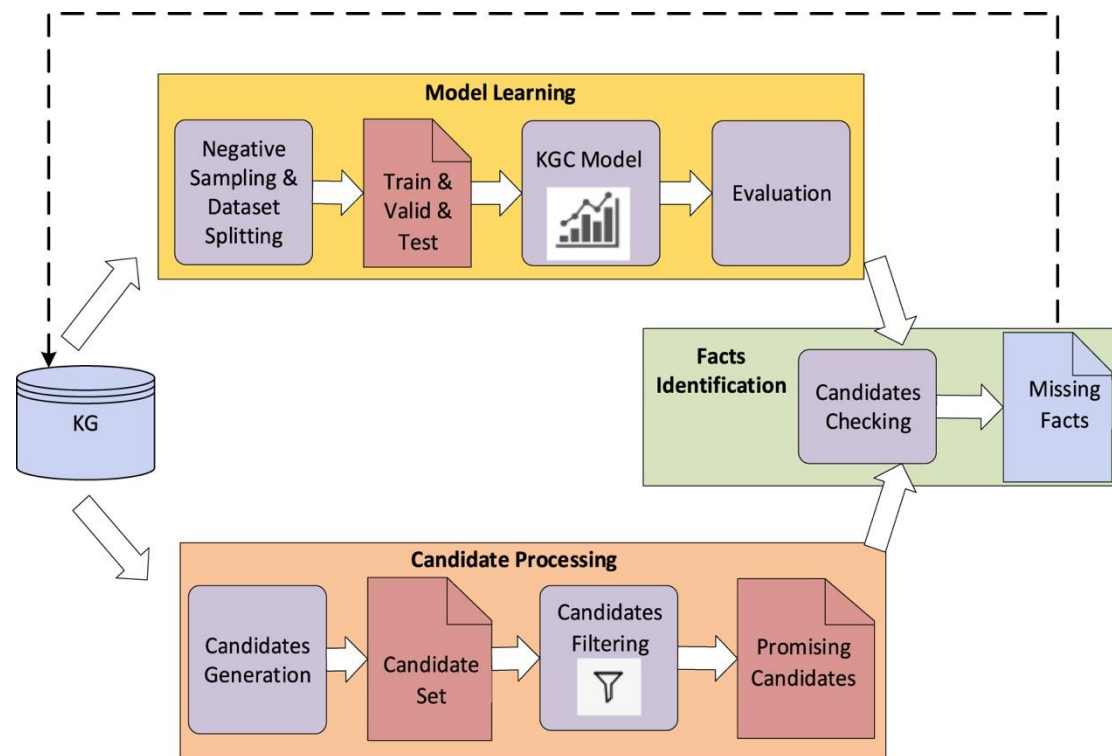


KNOWLEDGE GRAPH COMPLETION

- La definizione della score function consente a KGC di eseguire tre task fondamentali per garantire la qualità dei KG
- SCORE FUNCTION: varia a seconda del modello di KGC, è associata ad ogni fatto e, attraverso un punteggio, ne definisce la probabilità che esso sia corretto

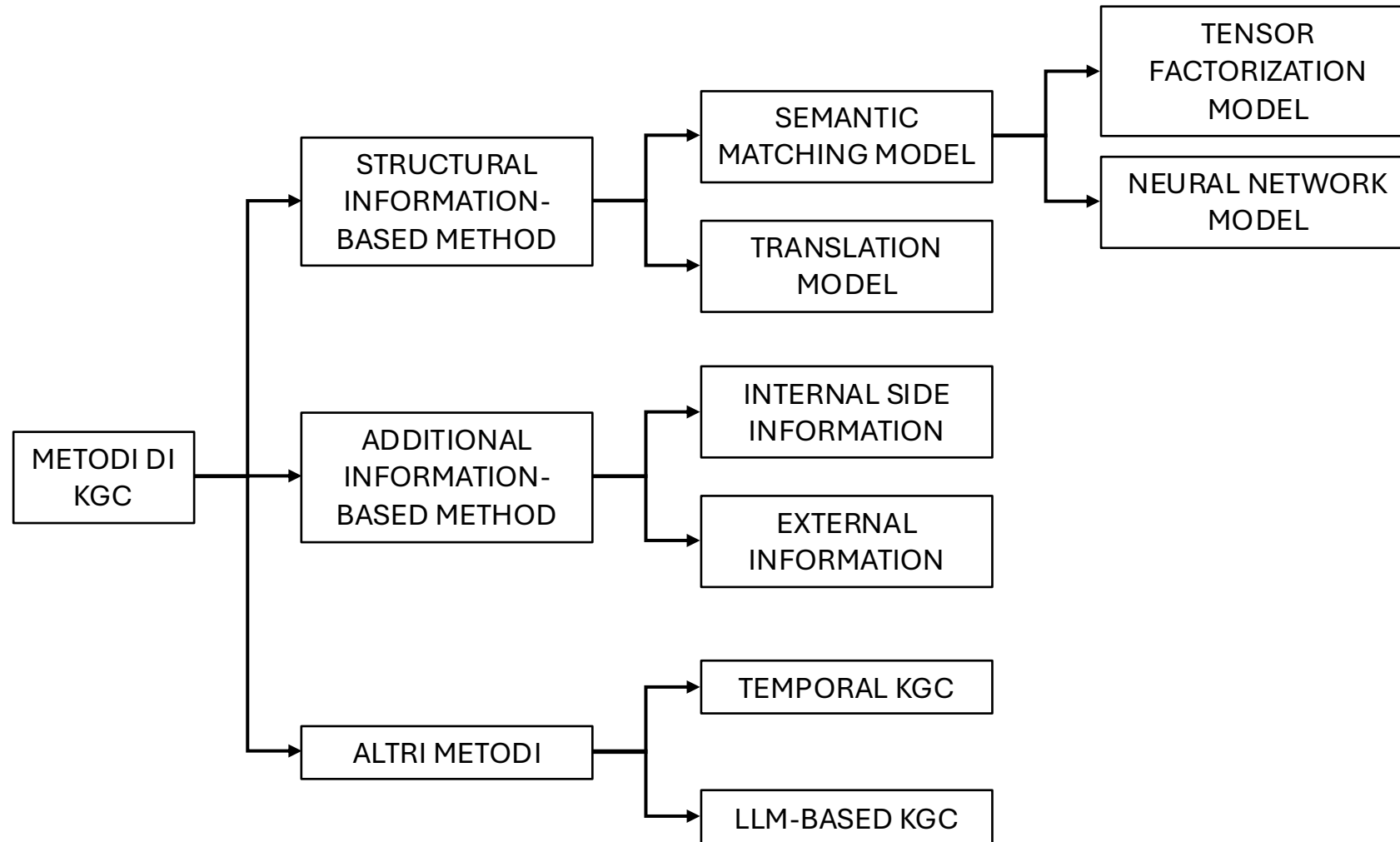


- TRIPLE CLASSIFICATION
- LINK PREDICTION
- RELATION PREDICTION





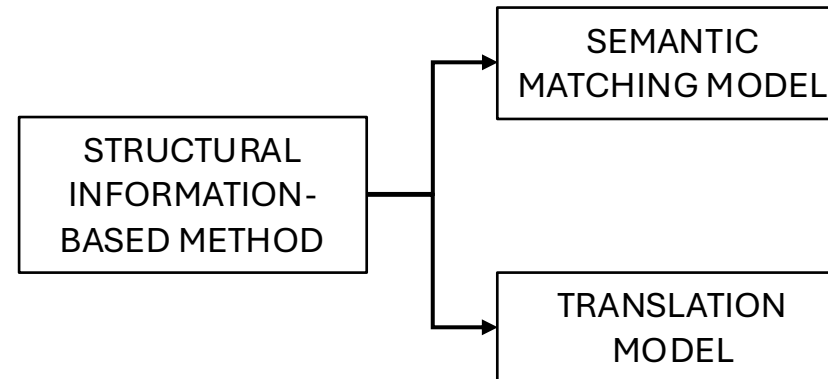
METODI DI KNOWLEDGE GRAPH COMPLETION

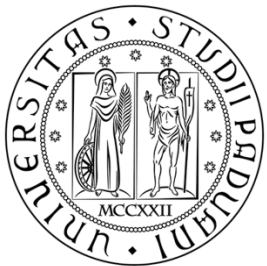




STRUCTURAL INFORMATION-BASED METHOD

- I metodi appartenenti a questa categoria sono basati sulle informazioni associate alle triple e alla loro struttura
- Per questo motivo, i risultati dei metodi sono fortemente dipendenti dalla densità dei fatti contenuti nel KG





SEMANTIC MATCHING MODEL

- La score function si basa sulle somiglianze semantiche tra embedding di entità e relazioni nello spazio rappresentato dal modello

TENSOR FACTORIZATION MODEL	NEURAL NETWORK MODEL
• I KG sono rappresentati come tensori • Il task di KGC viene visto come un problema di completamento di un tensore binario di terzo ordine • TuckER è il modello da cui derivano RESCAL, DistMult e ComplEx e presenta prestazioni migliori rispetto a questi ultimi per quanto riguarda il task di link prediction	• Le reti neurali possono cogliere le caratteristiche semantiche di entità e relazioni • Modellare le relazioni semantiche tra entità porta all'apprendimento di embedding più precisi • ConvE, ConvKB e CapsE sono stati presentati in quest'ordine e rientrano tra gli approcci basati su reti neurali convoluzionali



TRANSLATION MODEL

- La score function di traslazione stima la correttezza di una tripla basandosi su punteggi legati alla distanza
- Questa famiglia di metodi permette di codificare le entità come embedding a dimensione ridotta e le relazioni come vettori di traslazione
- Il modello di riferimento in questa categoria è **TransE**, in questo metodo, se $(head, relation, tail)$ è vera, allora $tail$ dovrebbe avvicinarsi all'embedding di $head$ sommato al vettore associato a $relation$
- Da un lato il successo di questo metodo deriva dal fatto che le traslazioni sono in grado di rappresentare in modo efficace le relazioni gerarchiche, dall'altro il modello mostra un limite importante in quanto esso è in grado di modellare solamente relazioni 1-1
- Alcuni metodi nati successivamente che permettono di modellare relazioni più complesse sono **TransH**, **TransR** e il più complesso **RotatE**, quest'ultimo ottiene le prestazioni migliori relativamente al task di link prediction



ADDITIONAL INFORMATION-BASED METHOD

- I metodi appartenenti a questa categoria non sono basati solo sulle informazioni associate alle triple e alla loro struttura ma integrano informazioni aggiuntive come descrizioni, immagini, vincoli o regole logiche

INTERNAL SIDE INFORMATION	EXTERNAL INFORMATION
• L'informazione intrinseca contenuta all'interno dei KG può fare riferimento ad attributi, entità o relazioni • ENTITY-RELATED INFORMATION: permette di evitare di sprecare risorse predicendo entità incompatibili con le relative relazioni • RELATION-RELATED INFORMATION: ha l'obiettivo di mettere in luce pattern relazionali come simmetrie o proprietà gerarchiche	• I metodi basati sull'informazione esterna sfruttano regole logiche o testi provenienti da fonti terze • RULE-BASED METHOD: le regole logiche possono fornire informazioni molto utili per KGC, in particolare la tecnica del rule mining permette di costruire regole logiche a partire dai fatti osservati • THIRD-PARTY DATA SOURCE-BASED METHOD: utilizzano fonti di dati terze come gli articoli presenti in Wikipedia



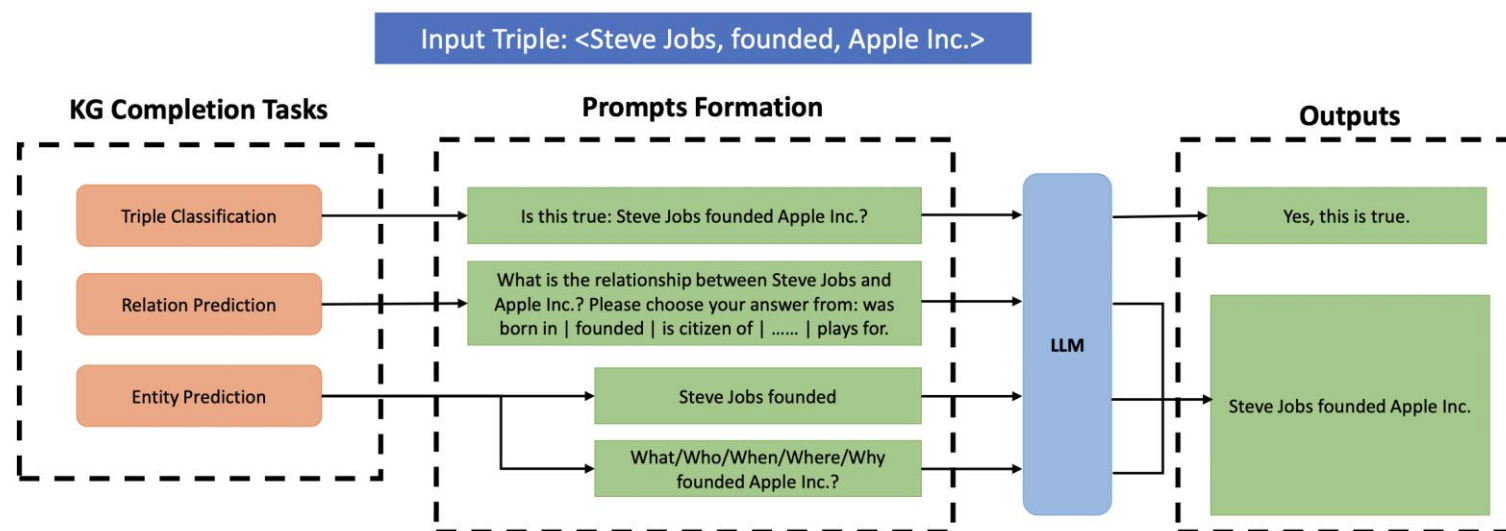
METODI DI TEMPORAL KNOWLEDGE GRAPH COMPLETION

- Buona parte dei fatti contenuti nei KG può risultare vera solamente in uno specifico istante o periodo temporale
- Al fine di risolvere questo problema assumono una rilevanza particolare i Temporal Knowledge Graph (TKG) sui quali si effettua Temporal Knowledge Graph Completion (TKGC)
- STRUTTURA: $(head, relation, tail, time)$ dove *time* può essere un istante o un intervallo temporale, ne deriva la definizione del task di time prediction
- Nella maggior parte dei casi, i metodi che includono l'informazione temporale nella score function sono caratterizzati da prestazioni migliori rispetto ai metodi classici
- Nel task di link prediction **TNTComplex** presenta risultati migliori rispetto a **TA-DistMult** e **TTransE**



LLM-BASED KNOWLEDGE GRAPH COMPLETION

- Gli LLM consentono di sfruttare la loro vasta conoscenza per massimizzare i risultati nel KGC
- Entità, relazioni e triple vengono trattate come sequenze di testo
- L'approccio KG-LLM effettua il tuning di LLM pre-addestrati
- È necessario convertire ogni richiesta in un prompt





DATASET



WordNet

A Lexical Database for English



KGC

DATASET	ENTITÀ	RELAZIONI	TRAINING	VALIDATION	TESTING
FB15k	14951	1345	483142	50000	59071
FB15k-237	14541	237	272115	17535	20466
WN18	40943	18	141442	5000	5000
WN18RR	40943	11	86835	3034	3134
YAGO3-10	123155	37	1078868	5000	5000

TEMPORAL KGC

DATASET	ENTITÀ	RELAZIONI	TRAINING	VALIDATION	TESTING	TIMESTAMP
ICEWS14	7128	230	72826	8941	8963	365
ICEWS05-15	10488	251	386962	46275	46092	4017



METRICHE DI VALUTAZIONE

- MEAN RECIPROCAL RANK: calcola la media del reciproco della posizione in classifica della risposta corretta

$$MRR = \frac{1}{Q} \sum_{i=1}^Q \frac{1}{rank_i}$$

- MEAN RANK: restituisce la media della posizione in classifica della risposta corretta

$$MR = \frac{1}{Q} \sum_{i=1}^Q rank_i$$

- HITS@N: indica, in percentuale, quante risposte corrette compaiono nelle migliori n posizioni in classifica

$$Hits@n = \frac{\text{count}(rank_i \leq n)}{Q}$$

- Q rappresenta il numero di triple da valutare e $rank_i$ indica la posizione della risposta corretta per la query i nella classifica generata grazie alla score function



APPLICAZIONI

- KG strutturati in modo corretto risultano utili in diverse applicazioni finali, ciò avviene anche grazie all'efficace acquisizione di conoscenza effettuata tramite KGC
- QUESTION ANSWERING: KGC garantisce la presenza delle informazioni necessarie per fornire una risposta derivante dai fatti contenuti nel KG a una domanda espressa nel linguaggio naturale
- RECOMMENDER SYSTEM: questa applicazione non sarebbe efficace senza KGC, i KG potrebbero risultare incompleti e ciò non permetterebbe di cogliere tutte le relazioni rilevanti compromettendo così la capacità di generare suggerimenti pertinenti
- DIALOGUE SYSTEM: hanno l'obiettivo di generare risposte fluenti e garantire una conversazione prolungata, i KG arricchiti tramite KGC sono in grado di apprendere al meglio il contesto del dialogo e predire con accuratezza le risposte più appropriate



CONCLUSIONE

- Le prestazioni dipendono in modo significativo dal dataset sul quale viene valutato ogni modello
- L'analisi di TKGC mette in evidenza l'importanza dell'integrazione di istanti o intervalli di tempo nei fatti che compongono il KG
- Gli sviluppi futuri nel campo del KGC comprenderanno l'integrazione tra metodi classici e Large Language Model