Paper Link: NELLIE: Never-Ending Linking for Linked Open Data | IEEE Journals & Magazine | IEEE Xplore

Title: NELLIE: Never-Ending Linking for Linked Open Data

ABSTRACT

The consistent generation of Linked Data-compliant Knowledge Graphs (KGs) remains a prevalent daily activity. However, the absence of comprehensive models for Linked Open Data (LOD) poses a significant obstacle to realizing the broader LOD vision. This paper introduces NELLIE, a meticulously engineered pipeline architecture designed to address this challenge. NELLIE's purpose is to facilitate the construction of an integrated knowledge graph from the LOD. The pipeline comprises a series of modules, each tailored to address specific data augmentation challenges. The initial step involves crawling existing knowledge graphs within the LOD cloud, identifying matching KG pairs, and employing a two-phase linking approach (ontology matching followed by instance matching) for each pair. Subsequently, NELLIE amalgamates these paired knowledge graphs into a unified entity. The resulting fused knowledge graph serves as an optimal data source for knowledge-driven applications such as search engines, question answering systems, digital assistants, and drug discovery initiatives. Evaluation of NELLIE demonstrates a notable enhancement, up to 94.44%, in the Hit@1 score for the link prediction task on the fused knowledge graph. Additionally, the two-phases linking approach exhibits a substantial runtime improvement compared to the estimated runtime of a naïve linking approach, underscoring the efficacy of NELLIE's methodology.

INTRODUCTION

People are making lots of different knowledge graphs that connect information online. These graphs are super useful for things like search engines, language processing, and finding recommendations. But here's the catch: there's no easy way to combine all these graphs into one big super useful graph. Right now, only a tiny fraction of these graphs are connected, making it hard to achieve the big vision of a fully connected web of knowledge. Connecting them manually is a huge chore, especially for giant graphs like DBpedia and Wikidata. Plus, the number and size of these graphs keep growing, and sometimes different groups publish the same information independently, leading to duplicates. To solve this, there are tools and methods to link and merge these graphs. One such tool is NELLIE, a system that brings together different knowledge graphs into one. It has three main parts: the core (the main engine), the application (where specific tasks are handled), and the publication (where the results are shared). NELLIE tackles challenges like matching similar parts in different graphs, linking them, merging them, and making a final combined graph. The goal is to have a 24/7 system that's always connecting and merging these knowledge graphs, kind of like a super-smart learner that never stops learning. The paper talks about how they built NELLIE, the methods they used, and how well it works for tasks like predicting links between different pieces of information. They acknowledge that fully testing NELLIE is tough right now but share the code for others to check out. The paper ends with a quick overview of what they discussed, what they found out, and what they plan to do next.

PRELIMINARY

In this section, we present the core of the formalization and notation necessary to implement NELLIE.

A. KNOWLEDGE GRAPH

A Knowledge Graph (KG) G is a set of triples $(s, p, o) \in (R \cup B) \times P \times (R \cup L \cup B)$, where R is the set of all resources, B is the set of all blank nodes, P is the set of all predicates, and L the set of all literals.

B. KNOWLEDGE GRAPHS MATCHING

Given a source KG G and a set of target KGs $T = \{G1 \cdots Gn\}$. The goal of KG matching is to rank all KGs within T based on their likelihood of containing entities that have the potential to be linked to entities in G.

C. LINK SPECIFICATIONS

Linking is based on the LIMES framework, using link specifications (LSs) with similarity measures (m) and operators (op). LSs express conditions for linking resources, and the WOMBAT algorithm is used to automatically generate LS.

D. ONTOLOGY MATCHING

Given two sets of classes Cs and Ct, the goal is to find pairs (ci, cj) such that a relation r (e.g., owl:equivalentClass) holds.

E. INSTANCE MATCHING

Given sets of resources Gs and Gt, the goal is to find pairs (s, t) such that a relation r (e.g., owl:sameAs) holds, producing a set of links or mappings with an optional similarity score.

F. KNOWLEDGE GRAPHS FUSION

The aim is to find a consolidated KG $Gs^{\oplus}t$ containing a fused version of related entities from source (Gs) and target (Gt) KGs. Fusion is performed based on a predefined fusion strategy operator $\oplus$.

G. LINK PREDICTION

Given a subset of true triples, the goal is to learn a scoring function φ for each possible triple. Linear models like TuckER, ComplEx, and DistMult are used, where the scoring function is either tensor factorization or a more complex neural network. The sigmoid function is applied for probability predictions.

The segments provide a comprehensive overview of key concepts in knowledge graphs, including their structure, matching, linking specifications, ontology and instance matching, fusion, and link prediction. The clear definitions and explanations make these complex concepts accessible. The inclusion of specific algorithms and models adds depth to understanding. Overall, the segments serve as a valuable reference for those exploring knowledge graph-related topics.

APPROACH

The NELLIE pipeline architecture consists of three layers: core, publication, and application. The core layer, highlighted in this paper, is crucial for the system's functionality. The core components include Knowledge Graph (KG) matching, linking, fusion, and embedding. The focus is on building a robust core layer that serves as the backbone of the system.

1. **KG Matching:**
    - Three methods for KG matching are implemented: Metadata-Based, Content-Based, and Manual KGs Matching.
    - Metadata-Based matching involves collecting KGs' metadata and using LIMES with exact match string similarity.
    - Content-Based matching retrieves text from literal objects, preprocesses it, and uses various similarity measures.
    - Manual KGs Matching involves manually selecting KGs for evaluation.
2. **Linking:**
    - Linking involves ontology matching followed by instance matching.
    - Class Matching is performed based on literal objects within classes, utilizing Content-Based Class Matching, LogMap, and FCA-Map.
    - Instance Matching focuses on owl:sameAs relation using LIMES, with a configurable threshold.
3. **Knowledge Graph Fusion:**
    - Fusion combines mappings of matched instances into a universal mapping.
    - An additive fusion operator is implemented, retaining all triples from the source KG and combining similar triples from the target KG.
    - Various fusion strategies are defined for literal objects of the same subject and predicate.
4. **KG Embedding & Link Prediction:**
    - Embedding models (TuckER, ComplEx, DistMult) are deployed for link prediction.
    - TuckER is based on Tucker decomposition, DistMult is a bilinear model, and ComplEx incorporates real and imaginary parts of embeddings.

EVALUATION & DISCUSSION

In this section, we evaluate each of the NELLIE components, i.e., KG matching, linking, fusion, and embedding, where we performed a set of experiments to evaluate the different techniques we implemented for each component.

The paper presents a detailed evaluation of three distinct approaches for knowledge graph (KG) matching and subsequent fusion. The assessment encompasses metadata-based matching, content-based matching, manual matching, and a two-phase linking process. Additionally, the impact of knowledge graph fusion on the link prediction task is explored in two scenarios. The methods are applied to biological domain KGs, and the study employs various similarity measures and frameworks for evaluation.

The paper's strength lies in its comprehensive evaluation, covering metadata-based, content-based, and manual KG matching, as well as the two-phase linking approach. This multi-faceted analysis provides a holistic understanding of the proposed techniques. The creation of a benchmark for content-based matching, considering diverse annotators and similarity measures, adds credibility to the evaluation.

The study showcases the effectiveness of content-based matching, achieving high F-Measure values using Jaccard and Cosine-TF-IDF measures. The comparison of BERT similarity with traditional measures provides valuable insights into the advanced language model's performance.

The manual KG matching in the biological domain contributes to the paper's rigor, ensuring a robust evaluation of subsequent components like ontology matching, instance matching, fusion, and link prediction. The inclusion of evaluation metrics, such as Pseudo-F-Measure, adds quantitative support to the assessment.

The two-phase linking approach's rationale is well-explained, emphasizing the reduction in overall runtime. The empirical demonstration of significant speedup further strengthens the paper's contribution.

In the fusion and link prediction task, the study introduces two scenarios, shedding light on the impact of KG fusion. The careful consideration of scenarios A and B, along with the use of various embedding models like TuckER, DistMulti, and ComplEx, adds depth to the evaluation. The discussion on potential reasons for performance variations, such as the nature of relations in KGs and hyper-parameter selection, showcases the paper's analytical depth.

While the paper demonstrates the positive impact of KG fusion on the link prediction task, it acknowledges the absence of a benchmark for such evaluation, emphasizing the need for future investigations and optimization of hyper-parameters.

In conclusion, this paper presents a thorough and insightful evaluation of KG matching and fusion techniques, contributing valuable findings to the field. The detailed methodologies, diverse evaluations, and thoughtful discussions make it a noteworthy contribution to the domain of knowledge graph research.

RELATED WORK

The paper introduces a novel concept of continuous, 24/7 linking over RDF knowledge graphs (KGs) for building a comprehensive model for the Linked Open Data (LOD). The related work is categorized into four research areas: knowledge graph matching, ontology/instance matching, data fusion, and knowledge graph embedding. The authors provide a brief overview of notable works in each area to establish the context for their own contributions.

1. **Knowledge Graph Matching:**
    - The paper identifies a gap in the literature regarding continuous linking over RDF KGs, emphasizing the novelty of their approach.
    - Knowledge graph matching techniques based on document similarity, specifically using topic modeling, are explored.
    - NELLIE adopts KG matching techniques based on document similarity, creating one document per KG. A comparison with Sleeman et al.'s method is provided, highlighting differences in document generation.
2. **Ontology and Instance Matching:**
    - ***Ontology Matching:***
        - The importance of ontology alignment for knowledge integration is emphasized. Several state-of-the-art systems, including LogMap, Codi, Chen et al.'s machine learning algorithm, and VeeAlign, are introduced.
        - LogMap is noted for its scalability and reasoning capabilities, while VeeAlign employs a deep learning-based model with dual-attention techniques.
    - ***Instance Matching:***
        - Declarative link discovery frameworks like SILK and LIMES are discussed for computing links between instances using property-based methods.
        - Various machine learning approaches for generating link specifications, such as Wombat and Eagle, are mentioned.
        - Specific instance matching methods, such as Serimi, Niu et al. 's semi-supervised learning algorithm, and Slint, are introduced, each addressing different aspects of the instance matching problem.
3. **Data Fusion:**
    - Data fusion's role in data integration, focusing on achieving data completeness and conciseness, is highlighted.
    - Challenges in data fusion, particularly dealing with uncertainty due to conflicting data values, are discussed.
    - The authors refer to and for surveys on data fusion techniques and challenges in knowledge graph fusion.
    - Linked data quality assessment and fusion framework Sieve is introduced as a tool based on the linked data integration framework (LDIF).
4. **Knowledge Graph Embedding:**
    - The paper acknowledges the significant development in Knowledge Graph Embedding (KGE) techniques for tasks such as graph completion, question answering, and link prediction.

    - Mentioned techniques include RESCAL, HolE, and TransE, each employing different approaches such as three-way factorization, circular correlation, and energy-based modeling.
    - The authors encourage further exploration of knowledge graph embedding approaches and applications, directing interested readers to.

The paper presents a well-structured overview of related work in key areas of knowledge graph research. The comprehensive coverage of literature and the contextualization of each research area contribute to the paper's coherence and provide readers with a solid foundation for understanding the proposed 24/7 linking approach. The paper effectively positions its contributions within the broader landscape of knowledge graph research.

CONCLUSION & FUTURE WORK

The paper presents a comprehensive pipeline architecture, NELLIE, for addressing challenges in knowledge graph augmentation and fusion. The incremental approach, with a focus on different stages of KG processing, is well-structured. The authors effectively communicate the ultimate goal of creating a continuously updated, fused knowledge graph and emphasize the ongoing nature of the project.

The implementation of various matching methods and integration with state-of-the-art systems demonstrates the practical aspects of NELLIE. The acknowledgment of the need for benchmarks and future plans for improvement, including the integration of additional approaches and fact-checking, adds credibility to the paper.

The paper could benefit from more explicit details on the implemented methods and strategies, providing readers with a clearer understanding of the technical aspects. Additionally, as the project progresses, regular updates on the achievements and challenges faced during the 24/7 implementation would enhance the transparency and reliability of NELLIE. Overall, the paper sets a solid foundation for the proposed architecture while leaving room for future advancements and refinements.