1 Introduction

1.1 Resumen:

This paper discusses the development of a graph query framework for relational learning, focusing on its logical and mathematical foundations. The existing approaches to relational learning are divided into two categories: connectionist and symbolic. While the connectionist approach has proven effective in many applications, the symbolic approach has faced challenges due to computational complexity arising from relational queries and a lack of robust frameworks for symbolic methods.

The proposed graph query framework addresses these issues by providing a query system with controlled complexity and stepwise pattern expansion capabilities using well-defined operations. This framework is especially suitable for use in relational machine learning as it allows for efficient extraction of characteristic relational patterns from data.

The paper's structure includes an overview of related research, introduction to the novel graph query framework with its main definitions and properties, representative query examples, analysis of computational complexity, implementation details for relational machine learning, and concluding remarks. The primary contribution of this work is the mathematical formalization of a graph query system that meets three key criteria: atomic operations, assessment of substructure beyond isolated nodes or complete graphs, and evaluation of cyclic patterns in polynomial time.

1.2 Evaluación:

Motivation:

Clarity: The introduction clearly explains the significance and relevance of the study, highlighting its importance in addressing the limitations of existing relational learning methods. It also effectively justifies the need for a novel graph query framework by discussing the challenges posed by computational complexity and lack of robust frameworks. The examples provided (e.g., social network analysis, protein characterization) help contextualize the problem and its potential impacts.

Improvement: To further strengthen the motivation, consider including more data or references to support the claims about the prevalence and importance of relational learning applications in different domains. This could help readers appreciate the breadth and significance of the problem being addressed by this research.

Novelty:

Originality: The introduction effectively describes the proposed approach's novelty, specifically emphasizing its unique ability to allow atomic operations for query expansion in a partitioned manner while maintaining polynomial-time complexity. It also highlights the lack of existing systems with these capabilities, positioning this work as an innovative contribution to the field.

Improvement: To more clearly differentiate this work from related efforts, consider explicitly comparing it with alternative graph query systems and discussing their limitations in addressing the issues identified in the introduction. This would help readers understand how the proposed framework advances the state of the art.

Clarity:

Comprehension: The introduction is well-written and easy to understand, using appropriate terminology and avoiding ambiguity. Complex ideas are presented clearly, and technical terms are defined when necessary. The structure of the paragraphs helps guide the reader through the different aspects of the problem and proposed solution.

Improvement: No specific improvements needed in this regard.

Grammar and Style:

Correctness: The introduction is free of grammatical and stylistic errors. It uses language appropriate for an academic setting, maintaining a professional tone throughout.

Improvement: No specific improvements needed in this regard.

Typos and Errors:

Accuracy: The introduction does not contain any typos or other errors that would affect comprehension or the credibility of the work.

Improvement: No specific improvements needed in this regard.

2 Related work

2.1 Resumen:

The related work section of the research paper discusses the common approach to executing relational queries, which involves developing patterns in an abstract representation of data and searching for their occurrences in actual datasets. This process falls under the scope of Graph Pattern Matching, a field that has been actively researched for over three decades. The paper distinguishes various methods of graph pattern matching based on structural, semantic, exact, inexact, optimal, approximate, isomorphic, graph simulation, and bounded simulation approaches.

The paper also highlights two fundamentally different types of relational learning models: the latent feature approach and the graph-pattern based approach. The latter focuses on automatically extracting relational patterns from data. Most pattern-based relational learning methods are derived from Inductive Logic Programming (ILP), which does not inherently offer relational classifiers but can be adapted for generating logical decision trees capable of managing relational predicates.

The paper reviews several algorithms and techniques related to pattern-based relational learning, including TILDE, MRDTL, Selection Graphs, DT-GBI, and Graph-Based Induction of Decision Trees. While some methods can learn complete graphs or node classifiers, the proposed technique in this paper supports learning from general subgraphs as base cases and can execute cyclic queries, enabling extraction of cyclic patterns during the learning process.

2.2 Evaluación:

Evaluation Levels:

Motivation:
Clarity: Does the section clearly explain the study's significance and relevance? Are the problem's importance and its wider impacts justified? (Provide specific examples from the text).

Improvement: The motivation is clear and well-justified. However, it could be strengthened by providing more concrete examples of how the proposed approach addresses existing limitations in related work. For instance, the authors could discuss how their method handles uncertainty or refines queries compared to other methods like TILDE or MRDTL.

Novelty:
Originality: Does the section clearly describe the proposed approach's novelty or originality? Does it differentiate itself from existing work? (Provide specific examples from the text).

Improvement: The novelty of the proposed approach is well-described, highlighting its ability to learn from general subgraphs and execute cyclic queries. However, the authors could further emphasize this by explicitly comparing their method with other approaches like TILDE, MRDTL, DT-GBI, which do not have these capabilities.

Clarity:

Comprehension: Is the section well-written and easy to understand? Does it use appropriate terminology and avoid ambiguity? (Provide specific examples from the text).

Correctness: Is the section free of grammatical and stylistic errors? Does it use language appropriate for an academic setting? (Provide specific examples from the text).

Improvement: The section is well-written and easy to understand. However, some parts could be rephrased for clarity. For instance, the sentence "Nevertheless, it does not cater to relational learning and therefore fails to offer certain operations for refining relational queries." could benefit from a more explicit explanation of what "certain operations" refers to. In addition, the use of abbreviations like "MRDTL" should be defined when they are first introduced in the text.

Grammar and Style:
Correctness: Is the section free of grammatical and stylistic errors? Does it use language appropriate for an academic setting? (Provide specific examples from the text).

Improvement: The grammar and style of the section are generally good. However, there is a minor error in "Another noteworthy pattern-based method for relational learning is Graph-Based Induction of Decision Trees (DT-GBI [16])" where "is" should be replaced with "are." Also, the sentence "As we have seen, some pattern-based approaches are able to learn to classify complete graphs, and some others construct node classifiers; our proposal supports learning from general subgraphs as base cases. Moreover, our technique can execute cyclic queries, hence allowing for extraction of cyclic patterns from data during the learning process." could be rephrased for better flow and clarity.

Typos and Errors:
Accuracy: Is the section free of typos and other errors? (Provide specific examples from the text).

Improvement: The section appears to be free of typos and other errors.

3 Relational machine learning

3.1 Resumen:

In the Relational machine learning section of a manuscript titled Logical-Mathematical Foundations of a Graph Query Framework for Relational Learning, the authors discuss using the graph query framework presented earlier to acquire relational classifiers on graph data sets. They explain how information gain pattern mining can be used to extract characteristic patterns from subgraph classes in the training set and how this process leads to the development of decision trees that separate the training set into distinct node classes.

The authors provide examples of applying this approach to simple social network graphs and more complex structures like character species classification in a Star Wars toy graph. They show that the relational decision trees generated by this method accurately assign types to nodes and extract meaningful patterns from the data, which can be used for direct assessment and future classification purposes.

3.2 Evaluación:

Evaluation Criteria: Must be Improved, Novelty, Clarity, Grammar and Style, Typos and Errors.

Motivation:
Clarity: The motivation section provides a clear explanation of the study's significance and relevance by highlighting the importance of graph data in various domains and emphasizing the need for efficient methods to classify subgraph patterns. However, it could be strengthened further by providing specific examples or statistics that demonstrate the prevalence of such problems in real-world applications.

Novelty: The section effectively describes the proposed approach's novelty by introducing the concept of relational machine learning and its application to graph query frameworks for subgraph classification. It also highlights how the proposed method differs from existing work, which typically employs binary decision trees. However, it could be improved by providing more explicit comparisons with related work and emphasizing the unique contributions of the proposed approach.

Comprehension: The section is well-written and easy to understand, using appropriate terminology and avoiding ambiguity. It effectively communicates complex ideas in a clear and concise manner. However, some sentences could be restructured for improved clarity, such as those introducing new concepts or describing technical details.

Correctness: The section is generally free of grammatical and stylistic errors. However, there are a few instances where the language could be tightened up to sound more precise and academic, such as using "the" instead of "a" when referring to specific examples.

Accuracy: There are no typos or other errors in this section.

4 Conclusions and future work

4.1 Resumen:

The paper's main contribution is a novel framework for graph queries that allows polynomial-time cyclic evaluation of queries and refinements based on atomic operations. This framework can be used in relational learning processes and fulfills several essential requirements, such as consistent grammar for both queries and evaluated structures, support for subgraph assessment beyond individual nodes, and effective refinement sets for top-down learning techniques.

Existing graph isomorphism-based query systems face exponential complexity issues when dealing with cyclic queries, but the proposed framework addresses this limitation by evaluating path and node existence/non-existence rather than demanding isomorphisms. The system has been implemented and experimentally demonstrated through a proof-of-concept implementation on a graph database and using the matplotlib library.

Despite its binary graph data set focus, the framework can be adapted to handle hypergraphs by maintaining path independence from edge arity. Future research should investigate more complex refinement families for specific learning tasks and develop automated methods for generating refinement sets based on graph dataset characteristics.

Additionally, patterns associated with leaf nodes in decision trees generated by the framework can characterize subgraph categories and justify decisions. The learned patterns can also serve as features in other machine learning methods like non-relational modeling. Furthermore, ensemble methods like random forests can utilize relational decision tree learning, and investigating probabilistic amalgamation of queries is essential to generate interpretable probabilistic decision tools.

Lastly, the paper suggests exploring additional machine learning algorithms alongside this query framework for further relational learning opportunities.

4.2 Evaluación:

Motivation:
Clarity: The section provides a clear explanation of the study's significance and relevance, highlighting the importance of graph query frameworks in various domains such as bioinformatics, cheminformatics, social network analysis, and computer vision. It also emphasizes the problem's wider impact on explainable learning and automatic feature extraction tasks. (Provided examples: "This is of great significance in both explainable learning and automatic feature extraction tasks."; "The results' graphs were obtained via our proof-of-concept implementation on a graph database and employing the matplotlib library [1].")

Improvement: The section could further strengthen its motivation by including more specific examples from these domains or by referencing relevant studies that have faced difficulties in handling cyclic queries. This would help to underscore the critical nature of the problem being addressed in this research.

Novelty:
Originality: The paper introduces a novel framework for graph queries that allows polynomial-time evaluation of cyclic patterns, distinguishing itself from existing graph isomorphism-based query systems. It also demonstrates the applicability of this framework to relational learning processes. (Provided examples: "The system offers a controlled and automated query construction via refinements, and the refinement sets constitute embedded partitions of the evaluated structure set, making them effective tools for top-down learning techniques."; "This work provides theoretical tools to support the accuracy of new refinement families. Future research will focus on developing automated methods to generate refinement sets based on a given learning task and the specific characteristics of the graph dataset.")

Improvement: The section could further emphasize the novelty by explicitly comparing the proposed approach with existing work, such as graph isomorphism-based query systems. This would help readers understand how the new framework addresses the limitations of these approaches.

Clarity:
Comprehension: The section is well-written and easy to understand, using appropriate terminology and avoiding ambiguity. It clearly explains the proposed approach's key concepts, such as query graph, refinement operation, and relational learning process. (Provided examples: "The concept of a path, which connects pairs of nodes, is independent of the edge arity involved."; "The system utilises a consistent grammar for both queries and evaluated structures.")

Improvement: No specific improvements are needed in this regard.

Grammar and Style:
Correctness: The section is free of grammatical and stylistic errors, using language appropriate for an academic setting.

Improvement: No specific improvements are needed in this regard.

Typos and Errors:
Accuracy: The section is free of typos and other errors.

Acknowledgements

Proyecto PID2019-109152G financiado por MCIN/AEI/10.13039/501100011033

DISARM project - Grant n. PDC2021-121197, and the HORUS project - Grant n. PID2021-126359OB-I00 funded by MCIN/AEI/310.13039/501100011033 and by the "European Union NextGenerationEU/PRTR"

References

References

Almagro-Blanco, P., & Sancho-Caparrini, F. (2017). Generalized graph pattern matching. *CoRR*, *abs/1708.03734*. URL: `http://arxiv.org/abs/1708.03734`. `arXiv:1708.03734`.

Barceló, P., Libkin, L., & Reutter, J. L. (2011). Querying graph patterns. In *Proceedings of the Thirtieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* PODS '11 (pp. 199–210). New York, NY, USA: ACM. URL: `http://doi.acm.org/10.1145/1989284.1989307`. doi:`10.1145/1989284.1989307`.

Blockeel, H., & Raedt, L. D. (1998). Top-down induction of first-order logical decision trees. *Artificial Intelligence*, *101*, 285–297. URL: `http://www.sciencedirect.com/science/article/pii/S0004370298000344`. doi:`10.1016/S0004-3702(98)00034-4`.

Bonifati, A. Fletcher, G.Voigt, H.Yakovets, N.Jagadish, H. V. (2018 Querying Graphs Morgan & Claypool Publishers

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 2787–2795). Curran Associates, Inc. URL: `http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf`.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? workforce implications. *Science*, *358*, 1530–1534.

Camacho, R., Pereira, M., Costa, V. S., Fonseca, N. A., Adriano, C., Simões, C. J., & Brito, R. M. (2011). A relational learning approach to structure-activity relationships in drug design toxicity studies. *Journal of integrative bioinformatics*, *8*, 176–194.

Chang, K.-W., Yih, S. W.-t., Yang, B., & Meek, C. (2014). Typed tensor decomposition of knowledge bases for relation extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. ACL – Association for Computational Linguistics. URL: `https://www.microsoft.com/en-us/research/publication/typed-tensor-decomposition-of-knowledge-bases-for-relation-extraction/`.

Consens, M. P., & Mendelzon, A. O. (1990). Graphlog: A visual formalism for real life recursion. In *Proceedings of the Ninth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* PODS '90 (pp. 404–416). New York, NY, USA: ACM. URL: `http://doi.acm.org/10.1145/298514.298591`. doi:`10.1145/298514.298591`.

Cook, S. A. (1971). The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing* STOC '71 (pp. 151–158). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/800157.805047. doi:10.1145/800157.805047.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., & Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '14 (pp. 601–610). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2623330.2623623. doi:10.1145/2623330.2623623.

Fan, W., Li, J., Ma, S., Tang, N., Wu, Y., & Wu, Y. (2010). Graph pattern matching: From intractable to polynomial time. *Proc. VLDB Endow.*, *3*, 264–275. URL: http://dx.doi.org/10.14778/1920841.1920878. doi:10.14778/1920841.1920878.

Gallagher, B. (2006). Matching structure and semantics: A survey on graph-based pattern matching. *AAAI FS*, *6*, 45–53.

García-Jiménez, B., Pons, T., Sanchis, A., & Valencia, A. (2014). Predicting protein relationships to human pathways through a relational learning approach based on simple sequence features. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*, 753–765. doi:10.1109/TCBB.2014.2318730.

Geamsakul, W., Matsuda, T., Yoshida, T., Motoda, H., & Washio, T. (2003). Classifier construction by graph-based induction for graph-structured data. In K.-Y. Whang, J. Jeon, K. Shim, & J. Srivastava (Eds.), *Advances in Knowledge Discovery and Data Mining: 7th Pacific-Asia Conference, PAKDD 2003, Seoul, Korea, April 30 – May 2, 2003 Proceedings* (pp. 52–62). Berlin, Heidelberg: Springer Berlin Heidelberg. URL: http://dx.doi.org/10.1007/3-540-36175-8_6. doi:10.1007/3-540-36175-8_6.

Gupta, S. (2015). *Neo4j Essentials*. Community experience distilled. Packt Publishing. URL: https://books.google.es/books?id=WJ7NBgAAQBAJ.

Luc De Raedt, Sebastijan Dumančić, Robin Manhaeve, Giuseppe Marra. *From Statistical Relational to Neural-Symbolic Artificial Intelligence.* In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI'20)*, 2021. ISBN: 9780999241165. Article No.: 688. Pages: 8. Yokohama, Japan.

Henzinger, M. R., Henzinger, T. A., & Kopke, P. W. (1995). Computing simulations on finite and infinite graphs. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on* (pp. 453–462). IEEE.

Jacob, Y., Denoyer, L., & Gallinari, P. (2014). Learning latent representations of nodes for classifying in heterogeneous social networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* WSDM '14 (pp. 373–382). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2556195.2556225. doi:10.1145/2556195.2556225.

Jiang, J. Q. (2011). Learning protein functions from bi-relational graph of proteins and function annotations. In T. M. Przytycka, & M.-F. Sagot (Eds.), *Algorithms in Bioinformatics* (pp. 128–138). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*, 3.

Karp, R. M. (1975). On the computational complexity of combinatorial problems. *Networks*, *5*, 45–68.

Knobbe, A. J., Siebes, A., Wallen, D. V. D., & Syllogic B., V. (1999). Multi-relational decision tree induction. In *In Proceedings of PKDD' 99, Prague, Czech Republic, Septembre* (pp. 378–383). Springer.

Latouche, P., & Rossi, F. (2015). Graphs in machine learning: an introduction.

Leiva, H. A., Gadia, S., & Dobbs, D. (2002). Mrdtl: A multi-relational decision tree learning algorithm. In *Proceedings of the 13th International Conference on Inductive Logic Programming (ILP 2003* (pp. 38–56). Springer-Verlag.

Milner, R. (1989). *Communication and Concurrency*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Nguyen, P. C., Ohara, K., Motoda, H., & Washio, T. (2005). Cl-gbi: A novel approach for extracting typical patterns from graph-structured data. In T. B. Ho, D. Cheung, & H. Liu (Eds.), *Advances in Knowledge Discovery and Data Mining: 9th Pacific-Asia Conference, PAKDD 2005, Hanoi, Vietnam, May 18-20, 2005. Proceedings* (pp. 639–649). Berlin, Heidelberg: Springer Berlin Heidelberg. URL: `http://dx.doi.org/10.1007/11430919_74`. doi:`10.1007/11430919_74`.

Wenfei Fan. *Graph Pattern Matching Revised for Social Network Analysis.* In *Proceedings of the 15th International Conference on Database Theory (ICDT '12)*, 2012. ISBN: 9781450307918. Publisher: Association for Computing Machinery. Address: New York, NY, USA. Pages: 8–21. Num. Pages: 14. Location: Berlin, Germany. DOI: 10.1145/2274576.2274578.

Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, Juncheng Liu, Bryan Hooi. *NodeAug: Semi-Supervised Node Classification with Data Augmentation.* En *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*, 2020, páginas 207–217. DOI: `https://doi.org/10.1145/3394486.3403063`.

Seyed Mehran Kazemi and David Poole. *RelNN: A Deep Neural Model for Relational Learning.* In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, University of British Columbia, Vancouver, Canada, 2018. Email: `smkazemi@cs.ubc.ca`, `poole@cs.ubc.ca`.

Maria Leonor Pacheco and Dan Goldwasser. *Modeling Content and Context with Deep Relational Learning. Transactions of the Association for Computational Linguistics*, 2021; 9, 100–119. DOI: `https://doi.org/10.1162/tacl_a_00357`.

K. Ahmed, A. Altaf, N.S.M. Jamail, F. Iqbal, R. Latif. *ADAL-NN: Anomaly Detection and Localization Using Deep Relational Learning in Distributed Systems. Applied Sciences*, 2023, 13, 7297. DOI: `https://doi.org/10.3390/app13127297`.

Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Maosong Sun. *Graph Neural Networks: A Review of Methods and Applications. AI open*, 2020, 1, 57-81.

Lingfei Wu, Peng Cui, Jian Pei, Liang Zhao, Xiaojie Guo. *Graph Neural Networks: Foundation, Frontiers and Applications*. En *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, 2022, páginas 4840–4841. DOI: `https://doi.org/10.1145/3534678.3542609`.

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, *104*, 11–33.

Namkyeong L., Dongmin H., Gyoung S. Na, Sungwon K., Junseok L. and Chanyoung P. (2023). Conditional Graph Information Bottleneck for Molecular Relational Learning.

Plotkin, G. (1972). Automatic methods of inductive inference .

van Rest, O., Hong, S., Kim, J., Meng, X., & Chafi, H. (2016). Pgql: a property graph query language. In *Proceedings of the Fourth International Workshop on Graph Data Management Experiences and Systems* (p. 7). ACM.

Reutter, J. L. (2013). *Graph Patterns: Structure, Query Answering and Applications in Schema Mappings and Formal Language Theory*. Ph.D. thesis Laboratory for Foundations of Computer Science School of Informatics University of Edinburgh.

Segaran, T., Evans, C., Taylor, J., Toby, S., Colin, E., & Jamie, T. (2009). *Programming the Semantic Web*. (1st ed.). O'Reilly Media, Inc.

Tang, L., & Liu, H. (2009). Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 817–826). ACM.

Zou, L., Chen, L., & Özsu, M. T. (2009). Distance-join: Pattern match query in a large graph database. *Proc. VLDB Endow.*, *2*, 886–897. URL: `http://dx.doi.org/10.14778/1687627.1687727`. doi:`10.14778/1687627.1687727`.

Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, Tianyu Wo. *Strong Simulation: Capturing Topology in Graph Pattern Matching. ACM Trans. Database Syst.*, January 2014, Volume 39, Number 1, Article No. 4. ISSN: 0362-5915. Publisher: Association for Computing Machinery. Address: New York, NY, USA. DOI: 10.1145/2528937. Keywords: dual simulation, graph simulation, data locality, Strong simulation, subgraph isomorphism.