Actes de la conférence BDA 2023, Montpellier

Reza Akbarinia, Tristan Allard, Angela Bonifati

BDA 2023

—

39ème Conférence sur la Gestion de Données
Principes Technologies et Applications

Reza Akbarinia[1], Tristan Allard[2], and Angela Bonifati[3]

[1]INRIA & LIRMM. reza.akbarinia@inria.fr
[2]Univ Rennes, CNRS, IRISA. tristan.allard@irisa.fr
[3]Université Lyon 1. angela.bonifati@univ-lyon1.fr

BDA 2023 : 39ème Conférence sur la Gestion de Données - Principes, Technologies et Applications

23-26 oct. 2023 Montpellier (France)

Actes de la conférence BDA 2023

Site web de la conférence : `https://bda2023.sciencesconf.org`

Table des matières

1 Message du Président et des Organisateurs **5**

2 Présidence & comités BDA 2023 **6**
2.1 Président des journées . . . 6
2.2 Comité d'organisation . . . 6
2.3 Comité de programme . . . 6
2.4 Comité de démonstration . . . 7
2.5 Comité du prix de thèse . . . 7

3 Conférences invitées **8**
3.1 Privacy-preserving publishing of Knowledge Graphs
Elena Ferrari (University of Insubria, Italie) . . . 8
3.2 Data Management for Deep Learning
Lei Chen (Hong Kong University of Science and Technology) . . . 8
3.3 Towards Practical, Scalable and Private Management of Cloud Data
Amr El Abbadi (University of California, Santa Barbara) . . . 9

4 Résumés des articles longs **10**
Evaluating Reification with Multi-valued Properties in a Knowledge Graph of Licensed Educational Resources . . . 14
Scalable Analytics on Multi-Streams Dynamic Graphs . . . 16
Parallel Pattern Enumeration in Large Graphs . . . 17
Computing Generic Abstractions from Application Datasets . . . 18
Scalable Reasoning on Document Stores via Instance-Aware Query Rewriting . . . 19
GPC : A Pattern Calculus for Property Graphs . . . 20
Mapping and Cleaning Open Commonsense Knowledge Bases with Generative Translation . . . 22
Evaluating the Factual Faithfulness of Graph-to-Text Generation . . . 24
Erebus : Explaining the Outputs of Data Streaming Queries . . . 25
An Analysis of Defects in Public JSON Schemas . . . 26
A Multidimensional Cost Model for Distributed Denormalized NoSQL Schemas . . . 27
Normalisations of Existential Rules : Not so Innocuous ! . . . 29
A Data-Driven Model Selection Approach to Spatio-Temporal Prediction . . . 30
Algorithms for Adaptive Upskilling with Learner Simulation . . . 32
Federated Learning on Personal Data Management Systems : Decentralized and Reliable Secure Aggregation Protocols . . . 35
Découverte de vérité confidentielle par calcul multi-parti . . . 37

Appliance Detection Using Very Low-Frequency Smart Meter Time Series . . . 38
ATEM : A Topic Evolution Model for the Detection of Emerging Topics in Scientific Archives . . . 40
ANTM : An Aligned Neural Topic Model for Exploring Evolving Topics . . . 42
Conjunctive Queries With Self-Joins, Towards a Fine-Grained Enumeration Complexity Analysis . . . 44
Efficient Enumeration of Recursive Plans in Transformation-based Query Optimizers . . . 45
Query Rewriting with Disjunctive Existential Rules and Mappings . . . 46

5 Résumés des articles courts 46
Efficient Computation of General Modules for ALC Ontologies . . . 48
Identification de données pertinentes dans des sources RDF . . . 50
Overview and Perspectives for Optimistic JSON Schema Witness Generation . . . 52
An End-to-End Machine Learning Framework for District Heating Networks Simulation . . . 53

6 Résumés des articles de démonstration 54
Computing MAP Inference on Temporal Knowledge Graphs with NeoMaPy . . . 56
More power to SPARQL : From paths to trees . . . 58
GSTSM Package : Finding Frequent Sequences in Constrained Space and Time . . . 59
InteGraal : a Tool for Data-Integration and Reasoning on Heterogeneous and Federated Sources . . . 61
PathWays : entity-focused exploration of heterogeneous data graphs . . . 62
Interpretable Clustering of Multivariate Time Series with Time2Feat . . . 63
Headwork : Powering the Crowd with Tuple Artifacts . . . 65
qEndpoint : A Wikidata SPARQL endpoint on commodity hardware . . . 67

7 Résumés des articles de doctorant 67
Representation Learning for Relational Structures . . . 69
Graph versioning for evolving urban data Versionnement de graphe pour les données urbaines évolutives . . . 70
Un graphe de données pour mesurer les Objectifs de Développement Durable et comprendre l'héritage des événements sportifs sur les villes . . . 72
Apprentissage Distribué à Grande Echelle des Embeddings d'Ontologies avec OWL2Vec . . . 74

Entropy Maximisation For Diverse Recommendations 76

8 Prix BDA 2023 **76**
8.1 Prix des articles de recherche 76
8.2 Prix des démonstrations 76
8.3 Prix des thèses en gestion de données 76

1 Message du Président et des Organisateurs

La conférence BDA : Gestion de Données – Principes, Technologies et Applications constitue le rendez-vous annuel incontournable de la communauté française de gestion de données. La 39ème édition de BDA, qui s'est déroulée du 23 au 26 octobre 2023 à Montpellier, a été un franc succès et nous sommes ravis d'avoir partagé ces journées avec vous.

Poursuivant une tradition riche et dynamique de rencontres annuelles de la communauté francophone en gestion de données, BDA 2023 a rassemblé les acteurs académiques et industriels de la recherche en gestion de données à soumettre leurs travaux récents pour présenter les défis et les avancées scientifiques dans ce domaine en pleine effervescence.

Avec 120 participants, BDA démontre une nouvelle fois le dynamisme de notre communauté ainsi que l'importance cruciale de la gestion de données dans le monde actuel. La recherche en gestion de données n'a jamais été aussi active, variée et ouverte sur d'autres champs de l'informatique. L'omniprésence des données massives transforme en profondeur notre société et pose de nombreux défis pour la communauté de recherche en informatique, tant sur le plan fondamental qu'appliqué. Cette évolution s'inscrit aujourd'hui dans le contexte de la science des données, avec un cercle vertueux entre les différentes communautés scientifiques concernés par la gestion, l'analyse et la valorisation de données volumineuses, hétérogènes, incomplètes, imprécises, produites dynamiquement et avec divers degrés de structuration. Les différentes phases du cycle de vie de ces données posent de nombreux défis pour la communauté de recherche en modélisation et gestion de données, tant sur le plan fondamental qu'appliqué. ,

Le programme scientifique de cette année était particulièrement riche avec 27 articles de recherche, dont 22 longs et 5 courts, 8 démonstrations et 5 articles de doctorants. Nous souhaitons à exprimer notre gratitude à tous les auteurs pour la qualité de leurs contributions et présentations. Nous avons également eu le plaisir d'accueillir trois conférenciers invités conférenciers invités, Elena Ferrari, Lei Chen et Amr El Abbadi, Elena Ferrari, Lei Chen et Amr El Abbadi, qui ont partagé leurs connaissances sur des sujets d'actualité. Nous les remercions pour l'excellence de leurs présentations qui ont éclairé la conférence.

Un des objectifs majeurs de BDA est d'offrir aux chercheuses et chercheurs, et en particulier aux doctorantes et doctorants, la possibilité de présenter leurs travaux à la communauté, y compris des travaux récents déjà publiés dans d'autres conférences. Les actes de la conférence proposent ainsi des résumés de toutes les contributions, publiées et non publiées, et sont complétés par une édition spéciale du journal *Transactions on Large-Scale Data and Knowledge-Centered Systems (TLDKS)* avec les versions étendues de quelques articles sélectionnés. Comme les années précédentes, BDA 2023 a également récompensé des contributions exceptionnelles, avec des prix décernés pour deux articles de recherche, une démonstration et deux thèses, reflétant lm'excellence et l'innovation de notre communauté.

Enfin, nous tenons à remercier surtout tous les membres de l'équipe d'organisation, ainsi que les membres des comités de programme, de démonstration, et du prix de thèse qui ont rendu possible cette nouvelle édition de BDA. Nous sommes également reconnaissants envers nos soutiens, notamment le Département Info du Lirmm, le Groupe de recherche AMIS de l'Université Paul-Valéry Montpellier 3, Inria, INSAVALOR et l'Université de Montpellier. Enfin, nos remerciements vont aux nombreux participants qui ont fait vivre cette remarquable édition 2023.

Bernd Amann, Président des journées
Reza Akbarinia, Président du Comité d'Organisation
Angela Bonifati, Président du Comité de Programme
Tristan Allard, Président du Comité de Démonstration

2 Présidence & comités BDA 2023

2.1 Président des journées

— Bernd Amann, LIP6, Université Pierre et Marie Curie

2.2 Comité d'organisation

— Reza Akbarinia, LIRMM, INRIA (président)
— Arnaud Castelltort, Université de Montpellier
— Elena Demchenko, LIRMM
— Virginie Feche, LIRMM
— Benoit Lange, INRIA
— Anne Laurent, Université de Montpellier
— Pierre Leroy, INRIA
— Florent Masseglia, INRIA
— Esther Pacitti, Université de Montpellier
— Pascal Poncelet, Université de Montpellier
— Maximilien Servajean, Université Paul Valéry
— Federico Ulliana, Université de Montpellier

2.3 Comité de programme

— Angela Bonifati, Université Lyon 1 (présidente)
— Nicolas Anciaux, INRIA
— Oana Balalau, Inria and École Polytechnique
— Ladjel Bellatreche, LIAS/ISAE-ENSMA
— Nofar Carmeli, INRIA
— Camelia Constantin, University Pierre et Marie Curie
— Stefania Dumbrava, ENSIIE
— Daniela Grigori, Laboratoire LAMSADE
— Francesco Guerra, University of Modena e Reggio Emilia
— Mirian Halfeld Ferrari, LIFO - Université d'Orléans
— Leonid Libkin, University of Edinburgh RelationalAI
— Silviu Maniu, Universite Paris-Saclay
— Zoltan Miklos, Université de Rennes 1
— Benjamin Nguyen, INSA Centre Val de Loire (FR)
— Noel Novelli, Aix-Marseille University
— Esther Pacitti, Univérsité de Montpelier
— Papotti Paolo, EURECOM
— Veronika Peralta, University of Tours
— Liat Peterfreund, ENS-Paris
— Claudia Roncancio, Grenoble INP
— Pierre Senellart, ENS
— Hala Skaf-Molli, University of Nantes
— Olivier Teste, IRIT

— Riccardo Tommasini, University of Tartu
— Farouk Toumani, UCA
— Nicolas Travers, Léonard de Vinci Pôle Universitaire
— Katerina Tzompanaki, CY Cergy Paris University
— Karine Zeitouni, Université de Versailles-St-Quentin

2.4 Comité de démonstration

— Tristan Allard, Université de Rennes (président)
— Peggy Cellier, IRISA
— Cedric du Mouza, CNAM
— Javier A. Espinosa-Oviedo, Université de Lyon
— Amélie Gheerbrant, Université de Paris
— David Gross-Amblard, Univ Rennes / Irisa Lab
— Shadi Ibrahim, Inria Rennes
— Zoubida Kedad, Universite de Versailles-St-Quentin
— Ioana Manolescu, Inria and Institut Polytechnique de Paris
— Jean-Marc Petit, LIRIS
— Iulian Sandu Popa, INRIA David Lab
— Shaoyi Yin, Paul Sabatier University

2.5 Comité du prix de thèse

— François Goasdoué, Université de Rennes (président)
— Bernd Amann, Sorbonne Université
— Nadine Cullot, Université de Bourgogne
— Frédérique Laforest, INSA Lyon
— Ioana Manolescu, INRIA
— Benjamin Nguyen, INSA Centre Val de Loire
— Marie-Christine Rousset, Université Grenoble Alpes
— Dan Vodislav, Cergy Paris Université

3 Conférences invitées

3.1 Privacy-preserving publishing of Knowledge Graphs
Elena Ferrari (University of Insubria, Italie)

Data sharing is crucial in the era of big data, and protecting users' sensitive information in these data is as vital as analyzing them. Knowledge graphs (KGs) are increasingly pivotal in data sharing due to their flexibility in modeling both attributes' values and relationships. However, due to the rich information in shared KGs, users' privacy is easier to breach. Thus, data providers must anonymize their KGs before sharing them. Unfortunately, data providers cannot straightforwardly use anonymization techniques developed for relational and traditional graphs to anonymize KGs as they do not consider both users' attributes and their relationships simultaneously. In this talk, we present a framework for anonymizing KGs, targeting three scenarios of increasing complexity : static publishing, sequential publishing, and personalized publishing. The first scenario allows data providers to publish their anonymized KGs once. The second one extends the first to enable the providers to publish new anonymized versions of their KGs. The final one lets users specify their privacy protection levels and anonymize KGs to protect all users under their levels.

Elena Ferrari is a professor of Computer Science at the University of Insubria (Italy), where she leads the STRICT SociaLab. She received her Ph.D. and M.Sc. degrees in Computer Science from the University of Milano (Italy). Her research interests are in the broad area of cybersecurity, privacy, and trust. Current research includes security and privacy for IoT, privacy-preserving data publishing, machine learning for cybersecurity, and blockchain. She is a fellow member of ACM and IEEE. She has been the recipient of several prestigious awards, including the 2009 IEEE Technical Achievement Award for pioneering contributions to Secure Data Management, the 2021 ACM SIGSAC Outstanding Contributions Award, the ACM CODASPY Research Award, and the ACM SACMAT 10-Year Test of Time Award. She is the recipient of the 2024 IEEE Innovation in Societal Infrastructure Award for pioneering and sustained contributions to the security and privacy of online social networks. In 2018, she was named one of the 50 most influential Italian women in tech.

3.2 Data Management for Deep Learning
Lei Chen (Hong Kong University of Science and Technology)

Deep learning (DL) has made significant progress and found wide application in various fields, like chaptGPT for question answering. However, the success and efficiency of DL models depend on proper data management. Training deep learning-based image classifiers is challenging without labeled data, and efficiency is hindered by large datasets, complex models, and numerous hyperparameters. Lack of validation and explanation limits model applicability. In this presentation, I will discuss three crucial issues in data management for deep learning : 1) effective data preparation for DL, including extraction, integration, and labeling ; 2) DL training optimization, involving data compression and computation graph optimization ; and 3) the importance of model explanation for robustness and transparency. I will conclude by highlighting future research directions.

Lei Chen is a chair professor in the data science and analytic thrust at HKUST (GZ), Fellow of the IEEE, and a Distinguished Member of the ACM. Currently, Prof. Chen serves as the dean of information hub, the director of Big Data Institute at HKUST, MOE/MSRA Information Technology Key Laboratory. Prof. Chen's research interests include Data-driven AI, knowledge graphs, blockchains, data privacy, crowdsourcing, spatial and temporal databases and query optimization on large graphs and probabilistic databases. He received his BS degree in computer science and engineering from Tianjin University, Tianjin, China, MA degree from Asian Institute of Technology, Bangkok, Thailand, and PhD in computer science from the University of Waterloo, Canada. Prof. Chen received the SIGMOD Test-of-Time Award in 2015, Best research paper award in VLDB 2022, .The system developed by Prof. Chen's team won the excellent demonstration award in VLDB 2014. Prof. Chen had served as VLDB 2019 PC Co-chair. Currently, Prof. Chen serves as Editor-in-chief of IEEE Transaction on Data

and Knowledge Engineering and an executive member of the VLDB endowment.

3.3 Towards Practical, Scalable and Private Management of Cloud Data
Amr El Abbadi (University of California, Santa Barbara)

Due to the widespread use of cloud applications, searching for data from a cloud server has become ubiquitous. However, accessing data stored in a cloud server comes with severe privacy concerns owing to numerous attacks and data breaches. Much research has focused on preserving the privacy of data stored in the cloud using various advanced cryptographic techniques. Our goal in this talk is to demonstrate how private access of data can become a practical reality in the near future. Our focus is on supporting oblivious queries and thus hide any associated access patterns on both private and public data. For private data, ORAM (Oblivious RAM) is one of the most popular approaches for supporting oblivious access to encrypted data. However, most existing ORAM datastores are not fault tolerant and hence an application may lose all of its data when failures occur. To achieve fault tolerance, we propose QuORAM, the first datastore to provide oblivious access and fault-tolerant data storage using a quorum-based replication protocol. For public data, PIR (Private Information Retrieval) is the main mechanism proposed in recent years. However, PIR requires the server to consider data as an array of elements and clients retrieve data using an index into the array. This requirement limits the use of PIR in many practical settings, especially for key-value stores, where the client may be interested in a particular key, but does not know the exact location of the data at the server. In this talk we will discuss recent efforts to overcome these limitations, using Fully Homomorphic Encryption (FHE), to improve the performance, scalability and expressiveness of privacy preserving queries of public data.

Amr El Abbadi is a Professor of Computer Science. He received his B. Eng. from Alexandria University, Egypt, and his Ph.D. from Cornell University. His research interests are in the fields of fault-tolerant distributed systems and databases, focusing recently on Cloud data management, blockchain based systems and privacy concerns. Prof. El Abbadi is an ACM Fellow, AAAS Fellow, and IEEE Fellow. He was Chair of the Computer Science Department at UCSB from 2007 to 2011. He served as Associate Graduate Dean at the University of California, Santa Barbara from 2021–2023. He has served as a journal editor for several database journals, including, The VLDB Journal, IEEE Transactions on Computers and The Computer Journal. He has been Program Chair for multiple database and distributed systems conferences, including most recently SIGMOD 2022. He currently serves on the executive committee of the IEEE Technical Committee on Data Engineering (TCDE) and was a board member of the VLDB Endowment from 2002 to 2008. In 2007, Prof. El Abbadi received the UCSB Senate Outstanding Mentorship Award for his excellence in mentoring graduate students. In 2013, his student, Sudipto Das received the SIGMOD Jim Gray Doctoral Dissertation Award. Prof. El Abbadi is also a co-recipient of the Test of Time Award at EDBT/ICDT 2015. He has published over 350 articles in databases and distributed systems and has supervised over 40 PhD students.

4 Résumés des articles longs

A Survey on SPARQL Query Relaxation under the Lens of RDF Reification

Ginwa Fakih
ginwa.fakih@univ-nantes.fr
LS2N, UMR 6004, Nantes Université
Nantes, France

Patricia Serrano-Alvarado
Patricia.Serrano-Alvarado@univ-nantes.fr
LS2N, UMR 6004, Nantes Université
Nantes, France

ABSTRACT

SPARQL query relaxation has been used to cope with the problem of queries that produce none or insufficient answers. The goal is to modify these queries to be able to produce alternative results close to those expected in the original query. Existing approaches generally relax the query constraints based on logical relaxations through RDFS entailment and RDFS ontologies. Techniques also exist that use the similarity of instances based on resource descriptions. These relaxation approaches defined for SPARQL queries over RDF triples have proved their efficiency. Nevertheless, significant challenges arise for query relaxation techniques in the presence of statement-level annotations, i.e., reification. In this survey, we overview query relaxation works with a particular focus on issues and challenges posed by representative reification models, namely, standard reification, named graphs, n-ary relations, singleton properties, and RDF-Star.

KEYWORDS

SPARQL query relaxation, statement-level annotations, reification models, query ranking, ontology-based relaxation, similarity of instances

ACM Reference Format:
Ginwa Fakih and Patricia Serrano-Alvarado. 2023. A Survey on SPARQL Query Relaxation under the Lens of RDF Reification. In *Proceedings of 39ème Conférence sur la Gestion de Données - Principes, Technologies et Applications (BDA 2023).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

This paper was submitted to the semantic web journal (SWJ) [2].

When a query evaluated over a dataset produces empty or insufficient answers, query issuers may try to modify the query constraints. This task is time-consuming and requires users with a profound knowledge of the data distribution and the schema of the dataset. An efficient way to cope with this problem was proposed in the domain of cooperative answering for deductive databases [3]. The original idea, is a relaxation method to expand the scope of the query dynamically by relaxing the constraints in the query. The goal is to generalize the user query to produce more answers.

BDA 2023, October 23–26, 2023, Montpellier, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

In this survey, we focus on SPARQL[1] queries over RDF[2] triples. SPARQL allows defining queries with triple patterns over which conjunctions, disjunctions, and optional patterns can be defined. Query issuers can use the OPTIONAL clause of SPARQL to specify which triple patterns can be ignored if they cannot be satisfied during the query evaluation. This idea is interesting but used to a limited extent because other forms of query relaxation can be used rather than simply dropping triple patterns defined as optionals. In particular, SPARQL query relaxation can be based on a logical relaxation of some of the query constraints by using RDFS entailment and RDFS ontologies. The major challenge of this approach, is that the number of relaxed queries grows combinatorially with the number of relaxation steps and the query size. Several approaches have been proposed to optimize the query relaxation task and generate relaxed queries efficiently. Mainly, the idea is to organize relaxed queries from the most specific to the most general and execute them in this order until obtaining *k-relevant* answers. But many relaxed queries may produce no new answers. To cope with this problem, the query execution ordering can be based on the *similarity of queries*. This similarity can be computed using *information content* which is based on the number of instances per class or property. Therefore the challenge is to identify the most similar relaxed queries that may produce new answers and to reduce the space of the relaxed queries that are executed until obtaining k answers.

Other query relaxation approaches use the similarity of instances based on resource descriptions. For example, similarity can be measured using an appropriate function that returns the distance between two attribute values. Such distance can be calculated differently, e.g., lexical similarities (like Jaccard similarity, Jensen-Shannon divergence, cosine similarity), overlap measures, page rank scores, etc. The similarity functions are data-dependent. Thus, the necessary number of functions depends on the different data types. Calculating the similarity of instances can be very costly. Further, multi-valued predicates (triples having a subject-predicate pair with several objects) may induce additional distances to compute. Frequently, the similarity of instances is done offline before the query processing. That is because the distances to compute can be significant, and calculating them dynamically during query processing can be unrealistic.

Moreover, statements about statements, also called statement-level annotations, are increasingly used. They allow specifying that a fact is true under a particular context. Context can concern temporal aspects, provenance, scores, weights, etc. *Reification* allows

[1]SPARQL 1.1 Query Language W3C Recommendation https://www.w3.org/TR/sparql11-query/
[2]https://www.w3.org/TR/rdf11-primer/

making statements about statements in a generic manner. For example, Figure 1 illustrates (a) a traditional triple that consists of two resources (nodes) related by a property relation (edge) and (b) a reified triple stating that Query_Language is a topic of an Educational_Resource "to some extent with a cosine similarity of 0.6". In a graph, reification can be seen as defining edges about edges. Reification can be done using several syntaxes that we call models. For example, there is standard reification [5], named graphs [1], singleton properties [6], RDF-star [4], etc.

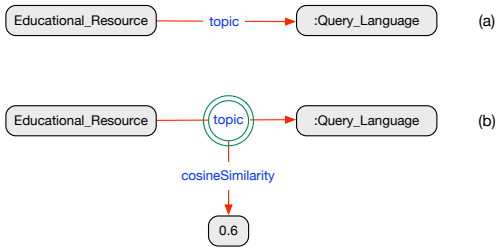


Figure 1: Representation of context on labeled edge

Existing query relaxation approaches have proved their efficiency but reification can lead to significant challenges for relaxation techniques.

The goal of this survey is to give a bird's-eye view of the SPARQL queries relaxation approaches over RDF datasets, and compare these approaches based on various relevant criteria.

2 CHALLENGES AND OPEN ISSUES

Applying query relaxation over queries whose constraints concern data and metadata (statement-level annotations) opens several issues. Current relaxation techniques are not proposed for querying metadata. (i) Logical relaxation over properties of metadata triple-patterns (with their superproperties) has in general no sense. (ii) Metadata values are not taken into account in the query ranking function that allows gathering the *k-relevant* answers closest to the original query. (iii) Current relaxation techniques are defined for equality of values but metadata constraints can include intervals of values in a range. (iv) Syntaxes to represent RDF reification and to query reified triples induce strange behavior in current query relaxation approaches.

Hence, we consider that new query relaxation contributions should be proposed to deal with queries querying data and metadata. A query rewritting process is necessary to distinguish the query constraints (which triple patterns concern data and which ones metadata) because relaxing metadata triple-patterns has not the same goal as relaxing other triple patterns. Both, ontology-based relaxation and similarity of instances, are necessary to query data and metadata. Depending on the application goals, these techniques may be combined. Querying data and metadata increases the query size which leads to the increase of the query relaxation lattice. Thus, optimal methods to prune this relaxation lattice should be used. Metadata values should play a role in the query ranking strategies that allow pruning the relaxation lattice but also gathering the *k-relevant* answers closest to the original query. Thus, new functions to calculate the similarity of queries should be proposed. Finally, relaxing metadata triple-patterns should take into account the type of metadata values. For the similarity of instances, the potential number of metadata types will be challenging. The number of similarity functions necessary to take into account all metadata types can be important which will add complexity to the similarity of instance approaches.

3 CONCLUSION

Applications querying reified triples may face the problem of empty or insufficient answers. Query relaxation approaches have been proposed to solve this problem but none of them is appropriate for metadata triple-patterns. In this paper, we provided an overview and comparative analysis of existing contributions focusing on SPARQL query relaxation. We also analysed and compared the syntaxes of some relevant reification models. Then, we underlined the potential effects of query relaxation approaches over reified triples. This survey has revealed that at the moment, no query relaxation solution deals with RDF triples and their annotations. Therefore, we pointed out some challenges and open issues in relaxing SPARQL queries under the lens of RDF reification, which we hope will open the doors to new inspiring contributions.

ACKNOWLEDGMENTS

This work has received a French government support granted to the Labex Cominlabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference ANR-10-LABX-07-01.

REFERENCES

[1] Jeremy J Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. 2005. Named graphs. *Journal of Web Semantics (JWS)* 3, 4 (2005), 247–267.
[2] Ginwa Fakih and Patricia Serrano-Alvarado. 2023. A Survey on SPARQL Query Relaxation under the Lens of RDF Reification. (2023). Submitted to Semantic Web Journal.
[3] Terry Gaasterland, Parke Godfrey, and Jack Minker. 1992. Relaxation as a platform for cooperative answering. *Journal of Intelligent Information Systems* 1, 3/4 (1992), 293–321.
[4] Olaf Hartig. 2017. Foundations of RDF* and SPARQL*:(An alternative approach to statement-level metadata in RDF). In *International Workshop on Foundations of Data Management and the Web (AMW)*, Vol. 1912. Juan Reutter, Divesh Srivastava.
[5] Frank Manola, Eric Miller, Brian McBride, et al. 2004. RDF primer. *W3C recommendation* 10, 1-107 (2004), 6.
[6] Vinh Nguyen, Olivier Bodenreider, and Amit Sheth. 2014. Don't like RDF reification? Making statements about statements using singleton property. In *Proceedings of the 23rd international conference on World Wide Web (WWW)*. 759–770.

Received 2 June 2023

Evaluating Reification with Multi-valued Properties in a Knowledge Graph of Licensed Educational Resources

Manoé Kieffer
manoe.kieffer@univ-nantes.fr
LS2N, UMR6004, Nantes Université
Nantes, France

Ginwa Fakih
ginwa.fakih@univ-nantes.fr
LS2N, UMR6004, Nantes Université
Nantes, France

Patricia Serrano Alvarado
patricia.serrano-alvarado@univ-nantes.fr
LS2N, UMR6004, Nantes Université
Nantes, France

ABSTRACT

This paper presents the construction of a Knowledge Graph (KG) of Educational Resources (ER) where RDF reification is essential. The ERs are described based on the subjects they cover considering their relevance. RDF reification is used to incorporate this subject's relevance. Multiple reification models with distinct syntax and performance implications for storage and query processing exist. This study aims to experimentally compare four statement-based reification models with four triplestores to determine the most pertinent choice for our KG. We built four versions of the KG. Each version has a distinct reification model, namely standard reification, singleton properties, named graphs, and RDF-star, which were obtained using RML mappings. The KG consists of 45,000 ERs, 13,000 authors, 135,000 subjects, and 8,250,000 relations linking the ERs to their subjects. Each of the four triplestores (Virtuoso, Jena, Oxigraph, and GraphDB) were setup four times (except for Virtuoso, which does not support RDF-star), and seven different SPARQL queries were experimentally evaluated. This study shows that standard reification and named graphs lead to good performance. It also shows that, in the particular context of the used KG, Virtuoso outperforms Jena, GraphDB, and Oxigraph in most queries. The recent specification of RDF-star and SPARQL-star sheds light on statement-level annotations. The empirical study reported in this paper contributes to the efforts towards the efficient usage of RDF reification. In addition, this paper shares the pipeline of the KG construction using standard semantic web technologies.

KEYWORDS

Knowledge graph, RDF reification, multi-valued properties, query evaluation, educational resources.

1 INTRODUCTION

When teachers want to create a new course, they typically do a keyword search for (open) Educational Resources (ER) on the web to reuse and integrate into their course. While there are numerous valuable and relevant resources available (such as slides, videos, figures, text, code, etc.), many remain undiscovered because they are not well connected.

ERs can be described by their title, authors, language, license, etc., as well as the subjects they cover. ERs' subjects can be numerous but not equally relevant for the ER. Some subjects are the main focus, while others are only mentioned briefly. Therefore, the relevance of each subject should be identified, and their relationship with each ER should be weighed accordingly. The best way to make ERs findable and reusable is to use the principles of the Linked Data. Semantic web technologies will allow a detailed description and interconnection of ERs. The recent specification of RDF-star and SPARQL-star sheds light on statement-level annotations. In fact, one of the first public work-drafts of RDF 1.2, introduces quoted triples as another kind of RDF term which can be used as the subject or object of another triple[1]. In our particular use case, statement-level reification will allow annotating with scores the relation of ERs and the subjects they treat. As the number of subjects can be important, this reified relation is a multi-valued property. Thus, efficiently dealing with multi-valued properties is important as well.

Multiple reification models with distinct syntax and performance implications for storage and query processing exist. The main objective of this work is to experimentally compare four statement-based reification models on four triplestores to determine the most pertinent choice for our KG.

The contributions of this paper are twofold: (i) a methodology to build four versions of a knowledge graph of ERs using statement-level reification, namely standard reification [1], singleton property [2], named graphs [3] and RDF-star [4], and (ii) an empirical evaluation of four triplestores (Virtuoso, Jena, GraphDB, and Oxigraph) with a set of seven SPARQL query templates grounded with up to six different instances (26 instantiated queries).

2 KNOWLEDGE GRAPH DESCRIPTION

Our project, aims to empower teachers to facilitate the creation of licensable ERs based on existing ones. To serve that purpose we created a KG of educational resources. The resources in our KG comprise unstructured ERs (documents, videos, and audio files, etc.), which are semantically annotated with DBpedia resources. By means of a wikification process, relevant DBpedia concepts related to ERs are used to provide a comprehensive description of each resource. Using RML-star [5] as a mapping language and Morph-KGC [6] as a mapper, we created one versions of our KG for each RDF reification model, namely standard reification, named graphs, singleton properties, RDF-star. The ontology of the four KGs can be seen in figure 1.

3 EXPERIMENTS

In order to test the impact of RDF reification models in combinaison with multi-valued properties we used the following methodology:

[1] https://w3c.github.io/rdf-concepts/spec/#section-triples

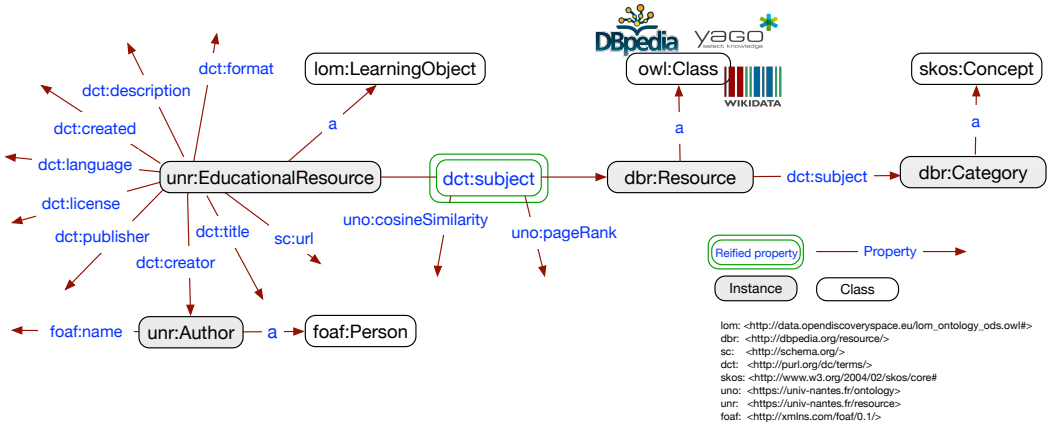


Figure 1: KG ontology.

(1) From JSON data we created four versions of our KG, respectivly using standard reification, singleton properties, named graphs, RDF-star.
(2) We instantiated one instance of Virtuoso, Jena, GraphDB, Oxigraph for each of the four different RDF reification model. Leading to a total of 15 instances (as Virtuoso doesn't support RDF-star).
(3) We created 7 query templates made to challenge the RDF reification model. We grounded the query template on 12 different instances chosen to challenge different multi-valued property sizes.
(4) For each instances, we ran each query three times in a row on a cold start.

Our final results show Virtuoso as being the most performant triplestore. And in term of RDF reification models, named graphs and standard reification end up being the best two models. Pairing Virtuoso and named graph or standard reification is the best approach for our specific KG.

4 CONCLUSION

This paper presented the pipeline for the generation of a knowledge graph (KG) of educational resources (ER) and the evaluation of several reification models with several triplestores. The objective was to identify the most suitable approach for this KG. To achieve this, we defined seven query templates instantiated in 26 grounded queries. Within the KG, reification was used in a multi-valued property to add two annotations whose range is between 0 and 1. Based on the insights derived from this experimental study, we were able to draw meaningful conclusions. Both, standard reification and named graphs with Virtuoso, exhibit similar performance. Named graphs show a slight advantage in some cases, in particular for join queries. RDF-star should be implemented more efficiently if quoted triples are included in RDF 1.2. Finally, for the KG presented in this paper, Virtuoso with named graphs, emerges as a good choice.

ACKNOWLEDGMENTS

This work has received a French government support granted to the Labex Cominlabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference ANR-10-LABX-07-01.
The authors thank Corentin Follenfant for his valuable contribution during the initial stages of this work. Additionally, the authors would like to acknowledge the Master's students in Computer Science at Nantes University for their involvement in various facets of this project.

REFERENCES

[1] Manola F, Miller E, McBride B, et al. RDF primer. W3C recommendation. 2004. Available from: https://www.w3.org/TR/rdf-primer/.
[2] Nguyen V, Bodenreider O, Sheth A. Don't like RDF reification? Making statements about statements using singleton property. In: International World Wide Web Conference (WWW); 2014. .
[3] Carroll JJ, Bizer C, Hayes P, Stickler P. Named graphs. Journal of Web Semantics. 2005.
[4] Hartig O. Foundations of RDF* and SPARQL* (An alternative approach to statement-level metadata in RDF). In: International Workshop on Foundations of Data Management and the Web (AMW); 2017. .
[5] Delva T, Arenas-Guerrero J, Iglesias-Molina A, Corcho O, Chaves-Fraga D, Dimou A. RML-star: A declarative mapping language for RDF-star generation. In: International Semantic Web Conference (ISWC) Posters, Demos and Industry tracks; 2021. .
[6] Arenas-Guerrero J, Iglesias-Molina A, Chaves-Fraga D, Garijo D, Corcho O, Dimou A. Morph-KGC star: Declarative generation of RDF-star graphs from heterogeneous data; 2022. Submitted to Semantic Web Journal (SWJ).

Scalable Analytics on Multi-Streams Dynamic Graphs

Angelos Anadiotis
Oracle
Lausanne, Switzerland

Muhammad Ghufran Khan
Inria & Institut Polytechnique de Paris
Palaiseau, France

Ioana Manolescu
Inria & Institut Polytechnique de Paris
Palaiseau, France

ABSTRACT

Several real-time applications rely on dynamic graphs to model and store data arriving from multiple streams. In addition to the high ingestion rate, the storage and query execution challenges are amplified in contexts where consistency should be considered when storing and querying the data. Our work addresses the challenges associated with multi-stream dynamic graph analytics. We propose a database design that can provide scalable storage and indexing, to support consistent read-only analytical queries (present and historical), in the presence of real-time dynamic graph updates that arrive continuously from multiple streams.

KEYWORDS

Dynamic graph, read-only present and historical queries, multi-stream graph processing

Parallel Pattern Enumeration in Large Graphs

Abir Farouzi1[1,3], Xiantian Zhou[2], Ladjel Bellatreche[1], Mimoun Malki[3], and Carlos Ordonez[2]

[1] LIAS/ISAE-ENSMA, France
[2] University of Houston, USA
[3] Ecole Nationale Supérieure en Informatique de Sidi Bel Abbès, Algeria

Abstract. Graphlet enumeration is a fundamental problem to discover interesting patterns hidden in graphs. It has many applications in science including Biology and Chemistry. In this paper, we present a novel approach to discover these patterns with queries, in a parallel database system. Our solution is based on an efficient partitioning strategy based on randomized vertex coloring, that guarantees perfect load balancing and accurate graphlet enumeration (complete and consistent). To the best of our knowledge, our work is the first to provide an abstract and efficient database solution with queries to enumerate both 3-vertex and 4-vertex patterns on large graphs.

Computing Generic Abstractions from Application Datasets

Nelly Barret
nelly.barret@inria.fr
Inria, IP Paris, France

Ioana Manolescu
ioana.manolescu@inria.fr
Inria, IP Paris, France

Prajna Upadhyay
prajna.u@hyderabad.bits-pilani.ac.in
BITS Pilani H, India

ABSTRACT

Digital data plays a central role in sciences, journalism, environment, digital humanities, etc. Open Data sharing initiatives lead to many large, interesting datasets being shared online. Some of these are RDF graphs, but other formats like CSV, relational, property graphs, JSON or XML documents are also frequent.

Practitioners need to *understand* a dataset to decide whether it is suited to their needs. Datasets may come with a schema and/or may be summarized, however the first is not always provided and the latter is often too technical for non-IT users. To overcome these limitations, we present an end-to-end *dataset abstraction* approach, which (*i*) applies on any (semi)structured data model; (*ii*) computes a *description* meant for human users, in the form of an Entity-Relationship diagram; (*iii*) integrates Information Extraction and data profiling to *classify* dataset entities among a large set of intelligible categories. We implemented our approach in a system called Abstra, and detail its performance on various datasets.

Scalable Reasoning on Document Stores via Instance-Aware Query Rewriting

Olivier Rodriguez, Federico Ulliana, and Marie-Laure Mugnier

LIRMM, Inria, Univ. Montpellier, CNRS Montpellier, France

Abstract. Data trees, typically encoded in JSON, are ubiquitous in data-driven applications. This ubiquity makes urgent the development of novel techniques for querying heterogeneous JSON data in a flexible manner. We propose a rule language for JSON, called constrained tree-rules, whose purpose is to provide a high-level unified view of heterogeneous JSON data and infer implicit information. As reasoning with constrained tree-rules is undecidable, we identify a relevant subset featuring tractable query answering, for which we design an automata-based query rewriting algorithm. Our approach consists of leveraging NoSQL document stores by means of a novel instance-aware query-rewriting technique. We present an extensive experimental analysis on large collections of several million JSON records. Our results show the importance of instance-aware rewriting as well as the efficiency and scalability of our approach.

GPC: A Pattern Calculus for Property Graphs

Nadime Francis
LIGM, Université Gustave Eiffel, CNRS
Champs-sur-Marne, France
nadime.francis@univ-eiffel.fr

Amélie Gheerbrant
Université Paris Cité, CNRS, IRIF
Paris, France
amelie@irif.fr

Paolo Guagliardo
University of Edinburgh
Edinburgh, UK
paolo.guagliardo@ed.ac.uk

Leonid Libkin*
University of Edinburgh
Edinburgh, UK
RelationalAI
Paris, France
l@libk.in

Victor Marsault
LIGM, Université Gustave Eiffel, CNRS
Champs-sur-Marne, France
victor.marsault@univ-eiffel.fr

Wim Martens
University of Bayreuth
Bayreuth, Germany
wim.martens@uni-bayreuth.de

Filip Murlak
University of Warsaw
Warsaw, Poland
f.murlak@uw.edu.pl

Liat Peterfreund
LIGM, Université Gustave Eiffel, CNRS
Champs-sur-Marne, France
liat.peterfreund@univ-eiffel.fr

Alexandra Rogova†
Université Paris Cité, CNRS, IRIF
Paris, France
rogova@irif.fr

Domagoj Vrgoč
PUC Chile & IMFD
Santiago de Chile, Chile
vrdomagoj@uc.cl

ABSTRACT

The development of practical query languages for graph databases runs well ahead of the underlying theory. The ISO committee in charge of database query languages is currently developing a new standard called *Graph Query Language* (GQL) as well as an extension of the SQL Standard for querying property graphs represented by a relational schema, called SQL/PGQ. The main component of both is the pattern matching facility, which is shared by the two standards. In many aspects, it goes well beyond RPQs, CRPQs, and similar queries on which the research community has focused for years.

Our main contribution is to distill the lengthy standard specification into a simple Graph Pattern Calculus (GPC) that reflects all the key pattern matching features of GQL and SQL/PGQ, and at the same time lends itself to rigorous theoretical investigation. We describe the syntax and semantics of GPC, along with the typing rules that ensure its expressions are well-defined, and state some basic properties of the language. With this paper we provide the community a tool to embark on a study of query languages that will soon be widely adopted by industry.

*Also with ENS, PSL University.
†Also with Data Intelligence Institute of Paris (diiP), Inria.

PODS '23, June 18–23, 2023, Seattle, WA, USA

ACM ISBN 979-8-4007-0127-6/23/06…$15.00
https://doi.org/10.1145/3584372.3588662

CCS CONCEPTS

• **Information systems → Graph-based database models**; **Query languages**.

KEYWORDS

graph databases, graph query languages, GQL, SQL/PGQ, pattern matching, syntax and semantics, expressive power, complexity, type systems

ACM Reference Format:
Nadime Francis, Amélie Gheerbrant, Paolo Guagliardo, Leonid Libkin, Victor Marsault, Wim Martens, Filip Murlak, Liat Peterfreund, Alexandra Rogova, and Domagoj Vrgoč. 2023. GPC: A Pattern Calculus for Property Graphs. In *Proceedings of the 42nd ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS '23), June 18–23, 2023, Seattle, WA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3584372.3588662

ACKNOWLEDGMENTS

This work was supported by the following grants: a Leverhulme Trust Research Fellowship; EPSRC grants N023056 and S003800; Agence Nationale de la Recherche projects ANR-21-CE48-0015 (Verigraph), ANR-18-CE40-0031 (QUID), ANR project EQUUS ANR-19-CE48-0019 which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 431183758; Poland's National Science Centre grant 2018/30/E/ST6/00042; ANID Millennium Science Initiative Program ICN17_002, and ANID Fondecyt regular project 1221799.

We thank the anonymous reviewers for their constructive feedback. We are also grateful to the members of ISO's SQL/GQL committee, and especially Fred Zemke, for their comments.

REFERENCES

[1] Renzo Angles, Marcelo Arenas, Pablo Barceló, Peter A. Boncz, George H. L. Fletcher, Claudio Gutierrez, Tobias Lindaaker, Marcus Paradies, Stefan Plantikow, Juan F. Sequeda, Oskar van Rest, and Hannes Voigt. 2018. G-CORE: A Core for Future Graph Query Languages. In *SIGMOD Conference*. ACM, 1421–1432.

[2] Renzo Angles, Marcelo Arenas, Pablo Barceló, Aidan Hogan, Juan L. Reutter, and Domagoj Vrgoc. 2017. Foundations of Modern Query Languages for Graph Databases. *ACM Comput. Surv.* 50, 5 (2017), 68:1–68:40.

[3] Pablo Barceló Baeza. 2013. Querying graph databases. In *PODS*. ACM, 175–188.

[4] Guillaume Bagan, Angela Bonifati, and Benoît Groz. 2020. A trichotomy for regular simple path queries on graphs. *J. Comput. Syst. Sci.* 108 (2020), 29–48.

[5] Pablo Barceló, Leonid Libkin, Anthony Widjaja Lin, and Peter T. Wood. 2012. Expressive languages for path queries over graph-structured data. *ACM Trans. Database Syst.* 37, 4 (2012), 31:1–31:46.

[6] Pablo Barceló, Jorge Pérez, and Juan L. Reutter. 2012. Relative Expressiveness of Nested Regular Expressions. In *AMW (CEUR Workshop Proceedings, Vol. 866)*. CEUR-WS.org, 180–195.

[7] Mikolaj Bojanczyk, Claire David, Anca Muscholl, Thomas Schwentick, and Luc Segoufin. 2006. Two-variable logic on data trees and XML reasoning. In *PODS*. ACM, 10–19.

[8] Béla Bollobás. 2013. *Modern Graph Theory*. Vol. 184. Springer Science & Business Media.

[9] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Moshe Y. Vardi. 2000. Containment of Conjunctive Regular Path Queries with Inverse. In *KR*. Morgan Kaufmann, 176–185.

[10] Mariano P. Consens and Alberto O. Mendelzon. 1990. GraphLog: A Visual Formalism for Real Life Recursion. In *PODS*. ACM Press, 404–416.

[11] Marco Console, Paolo Guagliardo, Leonid Libkin, and Etienne Toussaint. 2020. Coping with Incomplete Data: Recent Advances. In *PODS*. ACM, 33–47.

[12] Isabel F. Cruz, Alberto O. Mendelzon, and Peter T. Wood. 1987. A Graphical Query Language Supporting Recursion. In *SIGMOD Conference*. ACM Press, 323–330.

[13] Alin Deutsch, Nadime Francis, Alastair Green, Keith Hare, Bei Li, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Wim Martens, Jan Michels, Filip Murlak, Stefan Plantikow, Petra Selmer, Oskar van Rest, Hannes Voigt, Domagoj Vrgoc, Mingxi Wu, and Fred Zemke. 2022. Graph Pattern Matching in GQL and SQL/PGQ. In *SIGMOD Conference*. ACM, 2246–2258.

[14] Alin Deutsch, Yu Xu, Mingxi Wu, and Victor E. Lee. 2020. Aggregation Support for Modern Graph Analytics in TigerGraph. In *SIGMOD Conference*. ACM, 377–392.

[15] Nadime Francis, Amélie Gheerbrant, Paolo Guagliardo, Leonid Libkin, Victor Marsault, Wim Martens, Filip Murlak, Liat Peterfreund, Alexandra Rogova, and Domagoj Vrgoč. 2023. A Researcher's Digest of GQL. In *ICDT (LIPIcs, Vol. 255)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 1:1–1:22.

[16] Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. 2018. Cypher: An Evolving Query Language for Property Graphs. In *SIGMOD Conference*. ACM, 1433–1445.

[17] Alastair Green, Paolo Guagliardo, and Leonid Libkin. 2021. *Property graphs and paths in GQL: Mathematical definitions.* Technical Reports TR-2021-01. Linked Data Benchmark Council (LDBC). https://doi.org/10.54285/ldbc.TZJP7279

[18] Leonid Libkin, Wim Martens, and Domagoj Vrgoc. 2016. Querying Graphs with Data. *J. ACM* 63, 2 (2016), 14:1–14:53.

[19] Wim Martens, Matthias Niewerth, and Tina Trautner. 2020. A Trichotomy for Regular Trail Queries. In *STACS (LIPIcs, Vol. 154)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 7:1–7:16.

[20] Alberto O. Mendelzon and Peter T. Wood. 1995. Finding Regular Simple Paths in Graph Databases. *SIAM J. Comput.* 24, 6 (1995), 1235–1258.

[21] openCypher. 2017. Cypher Query Language Reference, Version 9. https://github.com/opencypher/openCypher/blob/master/docs/openCypher9.pdf

[22] Juan L. Reutter, Miguel Romero, and Moshe Y. Vardi. 2017. Regular Queries on Graph Databases. *Theory Comput. Syst.* 61, 1 (2017), 31–83.

[23] Oskar van Rest, Sungpack Hong, Jinha Kim, Xuming Meng, and Hassan Chafi. 2016. PGQL: a property graph query language. In *GRADES*. ACM, 7.

[24] Wikipedia contributors. 2020. GQL Graph Query Language. https://en.wikipedia.org/wiki/GQL_Graph_Query_Language

[25] Peter T. Wood. 2012. Query languages for graph databases. *SIGMOD Rec.* 41, 1 (2012), 50–60.

Mapping and Cleaning Open Commonsense Knowledge Bases with Generative Translation

Julien Romero
julien.romero@telecom-sudparis.eu
Télécom SudParis, SAMOVAR, IP Paris
France

Simon Razniewski
Simon.Razniewski@de.bosch.com
Bosch Center for AI
Germany

ABSTRACT

Structured knowledge bases (KBs) are the backbone of many knowledge-intensive applications, and their automated construction has received considerable attention. In particular, open information extraction (OpenIE) is often used to induce structure from a text. However, although it allows high recall, the extracted knowledge tends to inherit noise from the sources and the OpenIE algorithm. Besides, OpenIE tuples contain an open-ended, non-canonicalized set of relations, making the extracted knowledge's downstream exploitation harder. In this paper, we study the problem of mapping an open KB into the fixed schema of an existing KB, specifically for the case of commonsense knowledge. We propose approaching the problem by *generative translation*, i.e., by training a language model to generate fixed-schema assertions from open ones. Experiments show that this approach occupies a sweet spot between traditional manual, rule-based, or classification-based canonicalization and purely generative KB construction like COMET. Moreover, it produces higher mapping accuracy than the former while avoiding the association-based noise of the latter. Code and data are available at julienromero.fr/data/GenT.

KEYWORDS

Open Knowledge Bases, Generative Language Models, Schema Matching

1 INTRODUCTION

Motivation and Problem. Open Information Extraction (OpenIE) automatically extracts knowledge from a text. The idea is to find explicit relationships, together with the subject and the object they link. For example, from the sentence "In nature, fish swim freely in the ocean.", OpenIE could extract the triple *(fish, swim in, the ocean)*. Here, the text explicitly mentions the subject, the predicate, and the object. Therefore, if one uses OpenIE to construct a knowledge base (we call it an Open Knowledge Base, open KB) from a longer text, one obtains many predicates, redundant statements, and ambiguity.

OpenIE is often used for commonsense knowledge base (CSKB) construction. Previous works such as TupleKB [4], Quasimodo [8, 9] or Ascent [5–7] use OpenIE to extract knowledge from different textual sources (textbooks, query logs, question-answering forums, search engines, or the Web), and then add additional steps to clean and normalize the obtained data. Another example is ReVerb [1], which was used to get OpenIE triples from a Web crawl. The output of OpenIE typically inherits noise from sources and extraction, and the resulting KBs contain an open-ended set of predicates. This generally is not the case for knowledge bases with a predefined schema. Famous instances of this type are manually constructed, like ConceptNet [10] and ATOMIC [3]. They tend to have higher precision. Besides, they are frequently used in downstream applications such as question-answering [2, 12], knowledge-enhanced text generation [13], image classification [11], conversation recommender systems [15], or emotion detection [14]. These applications assume there are a few known predicates so that we can learn specialized parameters for each relation (a matrix or embeddings with a graph neural network). This is not the case for open KBs.

Still, many properties of open KBs, such as high recall and ease of construction, are desirable. In this paper, we study *how to transform an open KB into a KB with a predefined schema.* More specifically, we study the case of commonsense knowledge, where ConceptNet is by far the most popular resource. From an open KB, we want to generate a KB with the same relation names as ConceptNet. This way, we aim to increase precision and rank the statements better while keeping high recall. Notably, as we reduce the number of relations, we obtain the chance to make the statements corroborate. For example, *(fish, live in, water, freq:1)*, *(fish, swim in, water, freq:1)* and *(fish, breath in, water, freq:1)* can be transformed into *(fish, LocatedIn, water, freq:3)*, and therefore they all help to consolidate that statement. Besides, we make new KBs available to work with many existing applications originally developed for ConceptNet.

Transforming open triples to a predefined schema raises several challenges. In the simplest case, the subject and object are conserved, and we only need to predict the correct predefined predicate. This would be a classification task. For example, *(fish, live in, water)* can be mapped to *(fish, LocatedAt, water)* in ConceptNet. We could proceed similarly in cases where subject and object are inverted, like mapping *(ocean, contain, fish)* to *(fish, LocatedAt, ocean)*, with just an order detection step. However, in many cases, the object is not expressed in the same way or only partially: *(fish, live in, the ocean)* can be mapped to *(fish, LocatedAt, ocean)*. In other cases, part or all of the predicate is in the object, like *(fish, swim in, the ocean)* that can be mapped to *(fish, CapableOf, swim in the ocean)*. Here, the initial triple could also be mapped to *(fish, LocatedAt, ocean)*, showing that the mapping is not always unique. Other problems also arise, like with (near) synonyms. For example, we might want to map *(fish, live in, sea)* to *(fish, LocatedAt, ocean)*.

Approach and Contribution. We propose to approach the mapping of an open KB to a predefined set of relations as a translation task. We start by automatically aligning triples from the source and target KB. Then, we use these alignments to finetune a generative language model (LM) on the translation task: Given a triple from

an open KB, the model produces one or several triples in the target schema. The generative nature of the LM allows it to adapt to the abovementioned problems while keeping a high faithfulness w.r.t. the source KB. Besides, we show that this improves the precision of the original KB and provides a better ranking for the statements while keeping a high recall.

Our contributions are:

(1) We define the problem of open KB mapping, delineating it from the more generic KB canonicalization and the more specific predicate classification.
(2) We propose a generative translation model based on pre-trained language models trained on automatically constructed training data.
(3) We experimentally verify the advantages of this method compared to traditional manual and rule-based mapping, classification, and purely generative methods like COMET.

REFERENCES

[1] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *EMNLP*.
[2] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. *EMNLP* (2020).
[3] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (Comet-)Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In *AAAI*.
[4] Bhavana Dalvi Mishra, Niket Tandon, and Peter Clark. 2017. Domain-targeted, high precision knowledge extraction. *TACL* (2017).
[5] Tuan-Phong Nguyen, Simon Razniewski, Julien Romero, and Gerhard Weikum. 2021. Refined Commonsense Knowledge from Large-Scale Web Contents. *arXiv* (2021).
[6] Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2021. Advanced semantics for commonsense knowledge extraction. In *WWW*.
[7] Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2021. Inside ASCENT: Exploring a Deep Commonsense Knowledge Base and its Usage in Question Answering. *ACL* (2021).
[8] Julien Romero and Simon Razniewski. 2020. Inside Quasimodo: Exploring Construction and Usage of Commonsense Knowledge. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3445–3448. https://doi.org/10.1145/3340531.3417416
[9] Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z. Pan, Archit Sakhadeo, and Gerhard Weikum. 2019. Commonsense Properties from Query Logs and Question Answering Forums. In *CIKM*.
[10] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*.
[11] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. 2020. Multi-Label Classification with Label Graph Superimposing. In *AAAI*.
[12] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In *NAACL*.
[13] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A Survey of Knowledge-Enhanced Text Generation. *ACM Comput. Surv.* (2022).
[14] Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In *EMNLP*. Hong Kong, China.
[15] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion. In *KDD*.

Evaluating the Factual Faithfulness of Graph-to-Text Generation

Kun Zhang
kun.zhang@inria.fr
Inria & Institut Polytechnique de Paris
Palaiseau, France

Oana Balalau
oana.balalau@inria.fr
Inria & Institut Polytechnique de Paris
Palaiseau, France

Ioana Manolescu
ioana.manolescu@inria.fr
Inria & Institut Polytechnique de Paris
Palaiseau, France

ABSTRACT

Graph-to-text (G2T) generation takes a graph as input and aims to generate a fluent and faithful textual representation of the information in the graph. The task has many applications, such as dialogue generation and question answering. In this work, we investigate to what extent the G2T generation problem is solved for previously studied datasets, and how proposed metrics perform when comparing generated texts. To help address their limitations, we propose a new metric that correctly identifies factual faithfulness, i.e., given a triple (subject, predicate, object), it decides if the triple is present in a generated text. We show that our metric `FactSpotter` achieves the highest correlation with human annotations on data correctness, data coverage, and relevance. In addition, `FactSpotter` can be used as a plug-in feature to improve the factual faithfulness of existing models. Finally, we investigate if existing G2T datasets are still challenging for state-of-the-art models. This paper is accepted in Findings of EMNLP 2023 [1] and our code is available online: https://github.com/guihuzhang/FactSpotter.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; **Knowledge representation and reasoning**; **Neural networks**.

KEYWORDS

Graph-To-Text, Knowledge Graph, Constrained Text Generation

REFERENCES

[1] Kun Zhang, Oana Balalau, and Ioana Manolescu. 2023. FactSpotter: Evaluating the Factual Faithfulness of Graph-to-Text Generation. In *Findings of the EMNLP 2023*. Association for Computational Linguistics.

Erebus: Explaining the Outputs of Data Streaming Queries

Dimitris Palyvos-Giannas[1], Katerina Tzompanaki[2], Marina Papatriantafilou[1], and Vincenzo Gulisano[1]

[1] Chalmers University of Technology Gothenburg, Sweden
[2] ETIS, CY Cergy-Paris University, ENSEA, CNRS, UMR8051 Cergy, France

Abstract. In data streaming, why-provenance explains why a given outcome is observed but offers no help in understanding why an expected outcome is missing. Explaining missing answers has been addressed in DBMSs, but the solutions are not directly applicable to the streaming setting, because of the extra challenges posed by limited storage and by the unbounded nature of data streams. With Erebus we tackle the unaddressed challenges behind explaining missing answers in streaming applications, for the first time. Our thorough evaluation on real data shows that Erebus can explain the (missing) answers with small overheads, both in low- and higher-end devices.

An Analysis of Defects in Public JSON Schemas

Claire Yannou-Medrala and Fabien Coelho

Centre de recherche en informatique, Mines Paris – PSL University France

Abstract. JSON is a simple de facto standard cross-language textual format used to represent, exchange and store data and documents in computer systems. JSON Schema is a description language, based on JSON, proposed to describe JSON types and validate JSON data. We investigate over 57,800 distinct public schemas for various defects through static analysis, and identify cases of mistyping, misplacing, misnaming, misspelling, misversioning and other miscellaneous issues. Over 60% of schemas are defective, allowing in the worst case unintended data to be validated. These findings suggest to make key changes to the current JSON Schema draft so as to limit potential issues. It also leads us to design JSON Model, an alternative compact and expressive JSON data structure description language.

A Multidimensional Cost Model for Distributed Denormalized NoSQL Schemas

Jihane Mali
ISEP
Paris, France
jihane.mali@isep.fr

Faten Atigui
CEDRIC, Conservatoire National des Arts et Métiers (CNAM)
Paris, France
faten.atigui@cnam.fr

Ahmed Azough
Léonard de Vinci Pôle Universitaire, Research Center
Paris La Défense, France
ahmed.azough@devinci.fr

Nicolas Travers
Léonard de Vinci Pôle Universitaire, Research Center
Paris La Défense, France
nicolas.travers@devinci.fr

Shohreh Ahvar
Nokia Networks
Massy, France
shohreh.ahvar@nokia.com

ABSTRACT

The complexity of database systems has increased significantly along with the continuous growth of data, forcing Information Systems (IS) administrators to constantly adapt their data models and carefully choose the best option(s) for storing and managing data.In this context, we propose an automatic global approach for leading data model's transformation process. This approach starts with the generation of all possible solutions. It then relies on a cost model that helps to compare these generated data models to finally choose the best one for the given use case. This cost model integrates both data model and queries cost. It also takes into consideration the environmental impact of a data model as well as its financial and its time cost. This work presents for the first time a multidimensional cost model encompassing time, environmental and financial constraints, which compares data models leading to the choice of the best one for a given use case and context. In addition, a simulation tool for data model's transformation and cost computation has been developed based on our approach.

CCS CONCEPTS

• **Information systems → Parallel and distributed DBMSs**; **Database performance evaluation**; **Entity relationship models**.

ACM Reference Format:
Jihane Mali, Faten Atigui, Ahmed Azough, Nicolas Travers, and Shohreh Ahvar. 2023. A Multidimensional Cost Model for Distributed Denormalized NoSQL Schemas. In *BDA'23, 39Ã¨me ConfÃ©rence sur la Gestion de DonnÃ©es â Principes, Technologies et Applications, October 23-26, 2023, Montpellier, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

BDA'23, October 23-26, 2023, Montpellier, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Data's explosion especially characterized by the 3V (Volume, Variety & Velocity) has opened up major research issues related to modeling, manipulating, and storing massive amounts of data [4].

The resulting so-called NoSQL systems correspond to *four families* of data structures: key-value oriented (KVO), wide-column oriented (CO), document oriented (DO), and graph oriented (GO). Furthermore, requirements are constantly growing thus, guaranteeing the efficiency and availability of information requires the ability to respond effectively to new demands and requests on information systems. Better restructuring of database schemas is needed to maintain the 3V promise.

Today, this transformation is often driven by subjective choices, which makes it hard to take all the factors into account. The main issue is to provide the best data model for a given IS usage. This question leads to the cost model estimation of a solution according to the data model, statistics and queries to compute [3]. This question is hardly tackled for NoSQL solutions and especially when choosing the target architecture and structure. Finally, towards a responsible consumption for ODD 12[1] the question recently arose a need to be addressed.

To tackle this issue, we previously proposed a data model transformation approach[1, 2] which aims at producing a set of data models providing choices instead of focusing on a dedicated solution which prevents any trade-off, and reduces the search space by taking into account the use case (set of queries).

In this paper, our main contributions are:

- A multidimensional cost model, which integrates three key dimensions: time, environmental and financial costs,
- An environmental and a global cost-driven data model choice to target the optimal data model(s),
- Advanced evaluation of environmental cost to compare impacts of data models.

2 MULTIDIMENSIONAL COST MODEL

Our data model's generation process allows to generate a set of data models (Left part of Figure 1). Besides conventional measures

[1]UN ODD 12: https://www.un.org/sustainabledevelopment/sustainable-consumption-production/

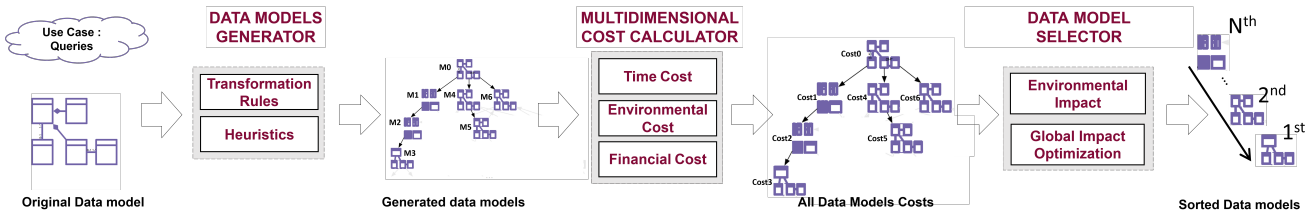


Figure 1: ModelDrivenGuide: Global Model Driven Approach

like response time and throughput, NoSQL database systems demand higher requirements due to the massive amount of data they need to handle (which incurs significant costs) in terms of storage, processing and communication.

In order to choose the best data model out of a set of generated possible solutions, we propose a cost model that automatically calculates the costs of data models to compare them and choose the best one. The main contribution of this work is to propose a multidimensional cost model based on **time** T, **environmental** E and **financial** F cost functions (as shown in Figure 1).

To achieve this we need to measure these subfunctions with common parameters: volumes (expressed in Bytes). Each cost dimension relies on the volume of stored data, the volume of processed data on servers and the volume of transferred data among servers. Moreover, each cost dimension combines those volumes in different ways, depending mostly on the data model itself and queries computation.

- **Time cost dimension**: varies according to several factors, including the size and complexity of the data model, the used storage type and processing infrastructure, and the speed of the network connection. It is expressed in seconds and calculated based on the volume of data processed by RAM access, storage access on SSD, and the volume of data transmitted.
- **Environmental cost dimension**: it is expressed in kg CO2e and depends on data accesses like RAM, storage and communication as well as number of servers (each server has a carbon footprint). Our aim is therefore to estimate carbon footprint (since measuring details of individual treatments is impossible) for the purposes of comparing data models, and not to obtain a precise impact.
- **Financial cost dimension**: it has few dependency on query execution, since most of the expenses come from the number of servers depending on the pricing model (*e.g.,* pay-as-you-go or subscription) and from external data transfers. Financial cost is expressed in currency (*e.g.,* €, $)

3 DATA MODEL SELECTOR

By considering the time, environment, and financial costs at the same level, we can make misleading decisions when evaluating data models. We should balance these factors based on their priorities, cost constraints, and sustainability goals to choose the most suitable data model that aligns with the use case.

The main goal of our approach is to choose an optimal data model. To achieve the choice, we can either focus on a given dimension or minimize the cost variation among settings.

- **Environmental impact optimization**: it focuses on the environmental dimension in order to target objectives from ODD 12. Under this policy, the environmental cost must be minimized while respecting the constraints of the information system, with guarantees of efficiency and budget. Out of all the generated data models, the goal is to find the data model(s) that minimizes a set of objectives $\{T, E, F\}$ in a given targeted setting (*e.g.,* data volume, #servers, etc.). This strategy minimizes E while checking if all queries' time are below the corresponding constraint and the budget is respected.
- **Global impact optimization**: To obtain a data model that minimizes all cost dimensions at once, we need to take into account its evolution on different parameters. In this way, the data model that varies least and has the lowest average cost is more likely to be retained over time. In order to measure this variation, we compute the cost density of each data model for all settings.

4 CONCLUSION

In this paper, we have proposed a multidimensional cost model that allows choosing the optimal data model out of different possible solutions. Our cost model integrates time, environmental and financial dimensions to define the cost of each data model. It considers queries cost on a data model and cost of data model itself.

We then propose a data model selector with two different approaches to choose the optimal data model(s). The first one chooses the data model that minimizes environmental cost w.r.t time and financial costs constraints, and the other one sorts data models according to their stability thanks to its costs' density among various settings.

REFERENCES

[1] Jihane Mali, Shohreh Ahvar, Faten Atigui, Ahmed Azough, and Nicolas Travers. 2022. A Global Model-Driven Denormalization Approach for Schema Migration. In International Conference on Research Challenges in Information Science. Springer, 529–545.
[2] Jihane Mali, Faten Atigui, Ahmed Azough, and Nicolas Travers. 2020. ModelDrivenGuide: An Approach for Implementing NoSQL Schemas. In DEXA'20. Springer, 141–151.
[3] M Tamer Özsu and Patrick Valduriez. 1999. Principles of distributed database systems. Vol. 2. Springer.
[4] Clarence JM Tauro, Shreeharsha Aravindh, and AB Shreeharsha. 2012. Comparative study of the new generation, agile, scalable, high performance NoSQL databases. International Journal of Computer Applications 48, 20 (2012), 1–4.

Normalisations of Existential Rules: Not so Innocuous!

David Carral
LIRMM, Inria, University of Montpellier, CNRS
Montpellier, France
david.carral@inria.fr

Lucas Larroque
DI ENS, ENS, CNRS, PSL University
Paris, France
lucas.larroque@ens.psl.eu

Marie-Laure Mugnier
LIRMM, Inria, University of Montpellier, CNRS
Montpellier, France
mugnier@lirmm.fr

Michaël Thomazo
Inria, DI ENS, ENS, CNRS, PSL University
Paris, France
michael.thomazo@inria.fr

ABSTRACT

Existential rules are an expressive knowledge representation language mainly developed to query data. In the literature, they are often supposed to be in some normal form that simplifies technical developments. For instance, a common assumption is that rule heads are atomic, *i.e.*, restricted to a single atom. Such assumptions are considered to be made without loss of generality as long as all sets of rules can be normalised while preserving entailment. However, an important question is whether the properties that ensure the decidability of reasoning are preserved as well. We provide a systematic study of the impact of these procedures on the different chase variants with respect to chase (non-)termination and FO-rewritability. This also leads us to study open problems related to chase termination of independent interest.

The full paper is available at https://hal-lirmm.ccsd.cnrs.fr/lirmm-03762686/file/carral-et-al-kr-2022.pdf, and the complete version with proofs at https://arxiv.org/abs/2206.03124

A Data-Driven Model Selection Approach to Spatio-Temporal Prediction*

Rocío Zorrilla[1], Eduardo Ogasawara[2],Patrick Valduriez[3], Fábio Porto[1]

[1]Laboratório Nacional de Computação Científica - LNCC

[2]Centro Federal de Educação Tecnológica Celso Sukow da Fonseca - CEFET-RJ

[3]INRIA & LIRMM

{romizc,fporto}@lncc.br, eogasawara@ieee.org, Patrick.Valduriez@inria.fr

Successfully predicting the behavior of spatio-temporal phenomena based on past observations is essential for a wide range of scientific studies and real-life applications like precipitation nowcasting [Souto et al., 2018], and climate alert systems [Murat et al., 2018]. In support of these applications, traditional data processing and time series analysis approaches generate predictive models that aim for predictive accuracy at the cost of high execution time and utilization of computational resources [Hassani and Silva, 2015].

More recently, a new class of systems, known as prediction serving systems, has emerged to support trained models scheduling warranting performance and run-time efficiency [Ghanta et al., 2019; Polyzotis et al., 2018]. For spatio-temporal phenomena, the focus of this paper, expressing a predictive query, involves specifying spatio-temporal constraints that define a region, a target variable whose values are to be inferred, and an evaluation metric for the performance of the predictive query. The query outcome then exhibits the target variable's future values on the specified region, computed by predictive models that meet the metric evaluation threshold.

However, we argue that building a query plan to answer a spatio-temporal predictive query is hard from several perspectives. Among them, we are interested in the model selection and allocation problem: for a given spatio-temporal query region, a serving system must automatically build an appropriate plan that chooses between training models or pick pre-trained models for each query region spatial position.

The main objective of this work was to develop an approach to make predictions, within some tolerated error margin, about future states of a spatio-temporal region, using carefully selected predictive models that have been trained with limited temporal data. To achieve this, we formulate the problem of model composition to process predictive queries and propose a solution where the model selection is guided by a data-driven approach backed by shape-based domain partitioning. The computational experiments were then designed to evaluate the proposal, considering the case study of temperature forecasting.

Within our proposal, both the domain partitioning (k–medoids) and the construction of Representative Models can be computed and persisted during an offline phase, quickly retrieved during an online phase, significantly reducing the elapsed time for processing predictive queries. In this regard, the choice of k becomes an important factor for the predictive quality, and three techniques to find optimal values of k were explored. We

*The authors thanks CAPES, CNPq, and FAPERJ for partially supporting the paper. This work is developed in the context of the HPDaSc INRIA-Brazil Associated Team.

find that the intuitive choice of a large value of k may not always produce the best results: fewer groups may produce more accurate results for some elements of the query region.

The previous result motivated the proposal of a neural network classifier for model selection. In the offline phase, we allow the construction of representative predictive models for multiple partitioning criteria ($k = \{8, 66, 132\}$). For the online phase, the classifier matches the subset (t_p time units) of each u.t.s in the query region to one of the representatives, thus creating the model composition for a given predictive query.

We show that our proposal can process predictive queries with significantly lower response time, while maintaining comparable predictive quality. To evaluate this experimentally, we used sMAPE forecast errors accumulated over query regions with MSE. Results indicate 20% and 45% relative increases for $k = 66$ and the Classifier approach, respectively, with a gain in computational efficiency of two orders of magnitude as a trade-off. We recognize that the Classifier needs to be improved, e.g., by considering a domain with a larger volume of data and understanding its classification accuracy.

Our proposal opens up several research directions. The calculation of pairwise DTW distances can be enhanced by grouping time series with an incremental process for the DTW matrix [Oregi et al., 2017]. For the domain partitioning task, we could consider non-crisp partitioning techniques [Izakian et al., 2015], producing more than one representative for a given element. This work did not focus on forecast time for the online phase as the ARIMA models deliver predictions in milliseconds; however, more complex models would imply significant service times. Therefore, a natural follow-up would include a multi-objective optimization process.

References

Ghanta, S., Subramanian, S., Khermosh, L., Sundararaman, S., Shah, H., Goldberg, Y., Roselli, D., and Talagala, N. (2019). ML health monitor: taking the pulse of machine learning algorithms in production. In *Applications of Machine Learning*, volume 11139, pages 191 – 202. International Society for Optics and Photonics, SPIE.

Hassani, H. and Silva, E. S. (2015). Forecasting with big data: A review. *Annals of Data Science*, 2(1):5–19.

Izakian, H., Pedrycz, W., and Jamal, I. (2015). Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, 39:235–244.

Murat, M., Malinowska, I., Gos, M., and Krzyszczak, J. (2018). Forecasting daily meteorological time series using arima and regression models. *International Agrophysics*, 32(2):253–264.

Oregi, I., Pérez, A., Del Ser, J., and Lozano, J. A. (2017). On-line dynamic time warping for streaming time series. In *Machine Learning and Knowledge Discovery in Databases*, pages 591–605, Cham. Springer International Publishing.

Polyzotis, N., Roy, S., Whang, S. E., and Zinkevich, M. (2018). Data lifecycle challenges in production machine learning: A survey. *SIGMOD Rec.*, 47(2):17–28.

Souto, Y. M., Porto, F., de Carvalho Moura, A. M., and Bezerra, E. (2018). A spatiotemporal ensemble approach to rainfall forecasting. In *IJCNN, 2018*, pages 1–8.

Algorithms for Adaptive Upskilling with Learner Simulation

Nassim Bouarour
CNRS, Univ. Grenoble Alpes
Grenoble, France
nassim.bouarour@univ-grenoble-alpes.fr

Idir Benouaret
Epita Research Laboratory
Lyon, France
idir.benouaret@epita.fr

Cédric d'Ham
CNRS, Univ. Grenoble Alpes
Grenoble, France
cedric.dham@univ-grenoble-alpes.fr

Sihem Amer-Yahia
CNRS, Univ. Grenoble Alpes
Grenoble, France
sihem.amer-yahia@univ-grenoble-alpes.fr

ABSTRACT

Upskilling is a fast-growing segment of the education economy [10]. Yet, there is little algorithmic work that focuses on crafting dedicated strategies to reach high skill mastery. In this paper, we formalize AdUp, an iterative upskilling problem that combines mastery learning [12] and Zone of Proximal Development [3]. We design two solutions for AdUp: MOO and MAB which adapt the difficulty of recommended tests to three objectives: learner's predicted performance, aptitude, and skill gap. Our simulation experiments, using two common learner simulation models: BKT (KT-IDEM) [11] and Item Response Theory (IRT) [13], demonstrate the necessity of leveraging all three objectives and the need to adapt the optimization objectives to the learner's progression ability.

1 INTRODUCTION

Today, learners engage in self-directed learning, managing many elements of their own study, which, in turn, often requires working on various learning activities independently with less direct guidance from teachers [7]. Consequently, providing guarantees on the quality of learning outcomes is increasingly difficult in these new bite-sized learning structures as they can lead to the so-called illusion of explanatory depth [14] where learners only acquire a superficial understanding of a topic. Ideally, each learner should receive tests chosen in a such way that the learner's skill progresses. This should account for the learner's ability to resolve tests based on skill and past performance. That is mastery learning [12].

Contributions. We formalize the AdUp Problem, our Adaptive Upskilling Problem as an optimization problem where a learner receives k tests that maximize expected performance and aptitude, and minimize accumulated skill gap. The combination of these objectives constitutes the novelty of our formalization. We propose to explore two solutions to solve AdUp: a Multi-Objective Optimization, referred to as MOO, and a Multi-Armed Bandits solution, referred to as MAB. MOO relies on Pareto dominance between k test sets and a *Hill Climbing* [9] heuristic algorithm that finds a subset of the non-dominated solutions [2]. Several variants can be drawn from MOO depending on the different compositions between the optimization objectives. We propose MAB, a second solution that chooses automatically which of the three optimization dimensions to optimize at each iteration. We formalize this approach as a multi-armed bandit (MAB) problem.

2 MODEL AND PROBLEM

We consider a learner $l \in \mathcal{L}$ who follows an iterative learning process for a skill sk. We focus on one skill that has a scalar as a value. At each step, l completes a set of k tests with different difficulty levels. Each test $t \in \mathcal{T}$ has a fixed difficulty d_t. We associate to each learner l a skill value $l.sk$ that either remains the same or increases as the learner successfully completes tests. To formalize our problem, we define dimensions that characterize the iterative learning process of a learner l for a skill sk.

2.1 Expected performance, aptitude, and gap

Expected performance. It is the expected performance of learner l for a test t. It is based on the similarity of t with successfully completed tests $l.\mathcal{S} \subseteq \mathcal{T}$ by l.

Aptitude. It quantifies the difference between a learner's skill value ($l.sk$) and the difficulty level of a test t (d_t). It represents the learner's progression ability for the skill when assigned tests that are correctly completed.

Gap. It quantifies the similarity between the past failed tests (set $l.\mathcal{F} \subseteq \mathcal{T}$) and the test t.

2.2 The AdUp problem

To achieve skill mastery, we propose an iterative formulation that solves the following optimization problem:

Problem 1 (The AdUp Problem). *Given a learner l, with a skill $l.sk$, find a batch $B \subseteq \mathcal{T}$ of k tests to assign to l at iteration i s.t.:*

$$\begin{aligned} &\textit{maximize} \sum_{t \in B} exPerf(l,t) \\ &\textit{maximize} \sum_{t \in B} apt(l,t) \\ &\textit{minimize} \sum_{t \in B} gap(l,t) \\ &\textit{subject to } |B| = k \end{aligned} \tag{1}$$

3 SOLUTIONS

The main challenge in solving AdUp, is its multi-objective nature. We propose to explore two solutions: a Multi-Objective Optimization, referred to as MOO, and a Multi-Armed Bandits solution, referred to as MAB.

3.1 Multi-Objective Optimization (MOO)

We propose an approach that finds the Pareto solutions by addressing all objectives at once [2]. To do so, we define a dominance relation between two sets of size k. We design a heuristic (Algorithm 1) to avoid an exhaustive exploration of the search space. It finds *times* optimal batches of tests using *Hill Climbing* to optimize expected performance and aptitude (Line 5). From the set of non-dominated candidates, the one with the lowest gap is chosen.

Algorithm 1: Heuristic MOO

```
   Input: learner l, set of tests T, size k, # repetition times
1  while not mastery do
2      Results ← ∅
3      for n in [1..times] do
4          C ← Random_candidate(k)
5          C* ← HCAE(C)
6          Results.Add(C*)
7      end
8      Keep non-dominated candidates in Results
9      B ← The solution from Results with the lowest skill gap
10     l completes B
11     l.sk ← skill_update(l.sk, B)
12 end
```

MOO **variants.** There are multiple solution variants to ADUP: MOO as described above; MOEG, MOAG, and MOAE optimize expected performance and gap, aptitude and gap, or aptitude and expected performance respectively; MOG, MOE, and MOA optimize gap only, expected performance only, or aptitude only respectively.

3.2 Multi-Armed Bandits Algorithm (MAB)

A drawback of the previous solution is that all the variants optimize exactly the same dimensions over all the assigned batches of tests during the whole learning process. However, it would be desirable to have an approach that can learn to find the dimensions to optimize at each iteration. We propose a MAB solution where each arm corresponds to one of the previous optimization variants and the reward r_i, at iteration i, for each variant v is defined as the speed of skill progression:

$$r_{iv} = \frac{\sum_{\forall \text{iterations } j, j<i} \text{skill gain offered by } v \text{ at iteration } j}{\#\text{time the variant } v \text{ was chosen}}$$

MAB **variants.** We implemented different multi-armed bandit strategies [15]: ϵ-GREEDY that chooses randomly a variant with an ϵ probability, THOMPSON Sampling which selects the arm with probability equal to the probability of it being the optimal choice, the upper confidence bound (UCB) which combines the reward and an uncertainty measure with a confidence degree (c) and SOFTMAX which relies on Boltzmann distribution with temperature (τ).

4 EXPERIMENTS

We formulate four research questions: **RQ1.** Is the combination of all optimization dimensions well-adapted for attaining mastery and improving skill gain? **RQ2.a** Do different settings of the skill update strategy exhibit different results? **RQ2.b** Does the choice of the learner simulation model impact mastery and skill gain? **RQ3.** Does an application of a meta-strategy that chooses to optimize a subset of dimensions at each iteration (MAB), improve mastery achievement?

4.1 Settings

Data. We use real data collected from a Czech educational system[1] from which we infer 42 distinct difficulty levels ranging in $]0, 1[$.
Learner simulation. There exist several models to simulate learners. The first model simulates learners using an extended version of BKT (KT-IDEM) [11] that considers the difficulty of tests. We used the implementation of [1]. The second model simulates learners based on latent factors [6]. It is based on Item Response Theory (IRT) [13]. We used the implementation of [4].
Baselines and Metrics. We compare MOO, MAB and their variants. We consider ALTERNATE, an approach that assigns a random set of tests whose difficulties alternate [8]. We report (1) the average skill gain, and (2) the average skill progression. We also examine (3) the percentage of learners who attained mastery and (4) the average number of iterations required to attain mastery.

4.2 RQ1: Impact of optimizing all dimensions

This experiment shows that combining all objectives yields the highest skill gain which permits a higher mastery in fewer iterations. It also shows challenging learners and optimizing aptitude is beneficial to attain mastery. These results also confirm the ZPD and Flow theories [3] and show the importance of leveraging aptitude and challenging learners.

4.3 RQ2.a: Impact of the skill update strategy

This experiment finds that MOO is not sensitive to varying different settings of the skill update strategy.

4.4 RQ2.b: Impact of changing the learner simulation model

This experiment finds that IRT generalizes the results of KT-IDEM. In this case, we also observe that MOO offers the highest rate of mastery. From the results, we see that the main difference between the two models is that IRT tends to favor gap as MOAG is comparable to MOO while BKT favors expected performance as MOAE was the second best.

4.5 RQ3: Impact of the meta-strategy

This experiment finds that choosing automatically the dimensions to optimize at each iteration improves the rate of mastery and the number of iterations needed to achieve it.

5 CONCLUSION AND FUTURE WORK

We addressed adaptive upskilling following a mastery learning approach. We proposed two approaches: MOO that directly solves our problem and a MAB that chooses among different optimization variants at each iteration. Our experiments confirmed that MAB offers a higher mastery rate and a better final skill gain than MOO.

For future work, we would like to deploy our algorithms with real learners. In addition to that, we aim to extend our formalization by considering additional theories (e.g., collaborative learning [5]).

[1] https://github.com/adaptive-learning/matmat-web/blob/master/data/data_description.md

REFERENCES

[1] Anirudhan Badrinath, Frederic Wang, and Zachary Pardos. 2021. pyBKT: An Accessible Python Library of Bayesian Knowledge Tracing Models. In *Proceedings of the 14th International Conference on Educational Data Mining*. 468–474.

[2] Ilaria Bartolini, Paolo Ciaccia, and Marco Patella. 2008. Efficient sort-based skyline evaluation. *ACM Transactions on Database Systems (TODS)* 33, 4 (2008), 1–49.

[3] Ashok R Basawapatna, Alexander Repenning, Kyu Han Koh, and Hilarie Nickerson. 2013. The zones of proximal flow: guiding students through a space of computational thinking skills and challenges. In *Proceedings of the ninth annual international ACM conference on International computing education research*. 67–74.

[4] bigdata ustc. 2021. EduCDM. https://github.com/bigdata-ustc/EduCDM.

[5] Kenneth A Bruffee. 1999. *Collaborative learning: Higher education, interdependence, and the authority of knowledge*. ERIC.

[6] Hao Cen, Kenneth Koedinger, and Brian Junker. 2006. Learning factors analysis–a general method for cognitive model evaluation and improvement. In *Intelligent Tutoring Systems: 8th International Conference, ITS 2006, Jhongli, Taiwan, June 26-30, 2006. Proceedings 8*. Springer, 164–175.

[7] Dragan Gašević, Vitomir Kovanović, Srećko Joksimović, and George Siemens. 2014. Where is research on massive open online courses headed? A data analysis of the MOOC Research Initiative. *International Review of Research in Open and Distributed Learning* 15, 5 (2014), 134–176.

[8] Masaki Matsubara, Ria Mae Borromeo, Sihem Amer-Yahia, and Atsuyuki Morishima. 2021. Task Assignment Strategies for Crowd Worker Ability Improvement. *Proc. ACM Hum. Comput. Interact.* 5, CSCW2 (2021), 1–20.

[9] Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, Pierre-Francois Dutot, and Denis Trystram. 2016. Multi-objective group discovery on the social web. In *ECML/PKDD*. Springer, 296–312.

[10] Higher Education Standards Panel. 2017. Final report–Improving retention, completion and success in higher education.

[11] Zachary A Pardos and Neil T Heffernan. 2011. KT-IDEM: Introducing item difficulty to the knowledge tracing model. In *International conference on user modeling, adaptation, and personalization*. Springer, 243–254.

[12] Radek Pelánek and Jiří Řihák. 2017. Experimental analysis of mastery learning criteria. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. 156–163.

[13] Mark D Reckase and Mark D Reckase. 2009. Unidimensional item response theory models. *Multidimensional item response theory* (2009), 11–55.

[14] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science* 26, 5 (2002), 521–562.

[15] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Federated Learning on Personal Data Management Systems: Decentralized and Reliable Secure Aggregation Protocols

Julien Mirval
julien.mirval@cozycloud.cc
Cozy Cloud, Inria-Saclay
UVSQ, Université Paris-Saclay
France

Luc Bouganim
luc.bouganim@inria.fr
Inria-Saclay
UVSQ, Université Paris-Saclay
France

Iulian Sandu-Popa
iulian.sandu-popa@uvsq.fr
UVSQ, Université Paris-Saclay
Inria-Saclay
France

ABSTRACT

The development and adoption of personal data management systems (PDMS) has been fueled by legal and technical means such as smart disclosure, data portability and data altruism. By using a PDMS, individuals can effortlessly gather and share data, generated directly by their devices or as a result of their interactions with companies or institutions. In this context, federated learning appears to be a very promising technology, but it requires secure, reliable, and scalable aggregation protocols to preserve user privacy and account for potential PDMS dropouts. Despite recent significant progress in secure aggregation for federated learning, we still lack a solution suitable for the fully decentralized PDMS context. This paper proposes a family of fully decentralized protocols that are scalable and reliable with respect to dropouts. We focus in particular on the reliability property which is key in a peer-to-peer system wherein aggregators are system nodes and are subject to dropouts in the same way as contributor nodes. We show that in a decentralized setting, reliability raises a tension between the potential completeness of the result and the aggregation cost. We then propose a set of strategies that deal with dropouts and offer different trade-offs between completeness and cost. We extensively evaluate the proposed protocols and show that they cover the design space allowing to favor completeness or cost in all settings.

CCS CONCEPTS

• **Computer systems organization → Peer-to-peer architectures**.

KEYWORDS

Secure aggregation, peer-to-peer, reliability, federated learning.

1 INTRODUCTION

New privacy-protection regulations (e.g., GDPR) and smart disclosure initiatives in the last decade have boosted the development and adoption of Personal Data Management Systems (PDMSs) [1]. A PDMS (e.g., Cozy Cloud [8], Nextcloud, Solid) is a data platform that allows users to easily collect, store, and manage into a single place data directly generated by the user's devices (e.g., quantified-self data, smart home data, photos) and data resulting from the user's interactions (e.g., social interaction data, health, bank, telecom). Users can then leverage the power of their PDMS to benefit from their personal data for their own good and for the benefit of the community [6].

As a result, the PDMS paradigm leads to a shift in the personal data ecosystem since data becomes massively distributed, on the user side. It also holds the promise of unlocking innovative usages. An individual can now cross her data from different data silos, e.g., health records and physical activity data. In addition, individuals can leverage their PDMSs by forming large communities of users sharing their data. This allows, for example, to compute statistics for epidemiological studies or to train a Machine Learning (ML) model for recommendation systems. In this context, it is natural to rely on a fully decentralized PDMS architecture (as opposed to central servers that raise several important issues such as cost, availability and scalability with the number of users), but this also poses new challenges.

Aggregation primitives are essential to compute basic statistics on user data and are also a fundamental building block for ML algorithms. In particular, Secure Aggregation (SA) is a central component of Federated Learning (FL), introduced in [12], as evidenced by the large body of recent work in this area [11]. However, to enable such new usages in the PDMS context, we need new solutions adapted to its specificity. First, PDMS users rely on large peer-to-peer systems for data sharing and computations [1, 5] thus requiring fully decentralized and scalable aggregation protocols, discarding data centralization on servers. Also, these protocols need to protect user privacy and adapt to varying selectivity (i.e., the consent of relevant participants). Ideally, the proposed protocol should provide an accurate result that takes advantage of the high-quality data available in PDMSs. Efficiency (i.e., protocol latency and total load of the system) is of prime importance given the potentially limited communication speed or computation power of PDMSs. Finally, given the scale of such decentralized aggregation, protocols must also be robust to node dropouts. To summarize, our goal is to design protocols that fulfill the following properties: **fully decentralized and highly scalable**, with the number of participants; **privacy-preserving**, i.e., protecting the confidentiality of the contributed user data; **accurate**, i.e., no trade-off between accuracy and privacy (e.g., like in the data anonymization or differential privacy approaches); **adaptable**, i.e., adapting to a large spectrum of computation selectivity values (reflecting the subset of contributor nodes) and system configurations (network and cryptographic latency); and **reliable**, i.e., handling node dropouts (e.g., failures, voluntary disconnections or unexpected communication delays).

Ensuring these properties altogether is challenging and to the best of our knowledge, the existing distributed Secure Aggregation (SA) protocols fail to achieve this objective. On one hand,

approaches such as local differential privacy are based on adding noise to protect privacy. This affects accuracy [3] or reliability to dropouts [15] and requires a very large number of participants to reduce the impact of noise which contradicts an adaptive node selectivity (see Section **??**). On the other hand, despite leveraging different cryptographic schemes in SA for FL [11] (e.g., encryption-based [2, 9] or secret sharing-based [4, 7, 10]), existing solutions employ a similar hybrid architecture wherein one or several highly available and powerful servers aggregate the data supplied by many user devices. Although some solutions consider the case of node dropouts, this applies to client devices and never to aggregation servers [4, 7]. In a Peer-to-Peer (P2P) PDMS system, all computations are performed by internal PDMS nodes (i.e., user devices). Hence, the data aggregators and data contributor nodes have the same constraints, i.e., limited computing power and availability. Such nodes cannot be expected to carry out heavy cryptographic operations [4] and can drop out during the computation. Fortunately, the P2P approach allows involving many nodes to perform a computation thus reducing the load on individual aggregators.

A first effort towards SA adapted to P2P systems was made in [13], where we designed a protocol that fulfill the above properties in an ideal setting, i.e., without considering the reliability issue. This work brings two major novelties. First, we focus on the reliability property, which is difficult to guarantee in a fully-decentralized setting and deserves a detailed study. Second, although our protocols apply to SA in general, we chose to study the more general case of FL, given its particular interest in the PDMS paradigm. The study of FL is also more challenging due to the potentially large size of the model, which increases the scalability problem. In our experiments, we consider model sizes from very small to very large, thus covering a wide range of use cases (including classical SA).

Our contributions are as follows. We analyze the impact of dropouts, be it contributor or aggregator nodes, on the other properties of an SA protocol designed for a P2P PDMS system. Node dropouts have a direct impact on accuracy (i.e., a single failure can make the final computation result useless) and on efficiency (i.e., it can introduce large latency). From this analysis, we derive the precise requirements of a reliable protocol and show that in a fully-decentralized context, reliability also introduces a tension between result completeness (i.e., the percentage of initial contribution in the final result, despite dropouts) and computation cost. We introduce the necessary building blocks to deal with these requirements. Then, we propose a variety of execution strategies offering different trade-offs between completeness and cost and allowing to cover a wide spectrum of dropout rates, contributor selectivity or trained model sizes. Our extensive experimental evaluation shows that the proposed strategies cover well the design space allowing to favor completeness or cost in all settings.

The full version of the paper [14] is structured as follows. We first discuss the related work w.r.t. the required properties. We then introduce the considered architecture and threat model. The next section reminds the main design principles proposed in [13] and then introduces, as a starting point, a straw-man SA protocol which efficiently computes the required aggregation assuming an ideal world (i.e., there are no node dropouts). This allows to highlight the challenges induced by reliability issues. We then present the necessary building blocks to addresses the reliability related challenges, before proposing four SA strategies that leverage those building blocks and allow for different trade-off between result completeness and aggregation cost. Finally, we extensively evaluate the proposed strategies and conclude.

ACKNOWLEDGMENTS

This work has been supported by the ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR.

REFERENCES

[1] Nicolas Anciaux, Philippe Bonnet, Luc Bouganim, Benjamin Nguyen, et al. 2019. Personal Data Management Systems: The Security and Functionality Standpoint. *Information Systems* (2019).
[2] Johes Bater, Gregory Elliott, Craig Eggen, Satyender Goel, et al. 2017. SMCQL: Secure Query Processing for Private Data Networks. *PVLDB* (2017).
[3] Aurélien Bellet, Rachid Guerraoui, Mahsa Taziki, and Marc Tommasi. 2018. Personalized and Private Peer-to-Peer Machine Learning. In *AIStat.*
[4] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, et al. 2017. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *ACM CCS.*
[5] Luc Bouganim, Julien Loudet, and Iulian Sandu Popa. 2023. Highly Distributed and Privacy-Preserving Queries on Personal Data Management Systems. *The VLDB Journal* (2023).
[6] EU Commission. 25 October 2020. Proposal for a Regulation on European Data Governance (Data Governance Act), COM/2020/767. [eur-lex].
[7] Henry Corrigan-Gibbs and Dan Boneh. 2017. Prio: Private, Robust, and Scalable Computation of Aggregate Statistics. In *NSDI.*
[8] Cozy Cloud. 2023. *Cozy Cloud (See https://cozy.io/fr/).*
[9] Ye Dong, Xiaojun Chen, Kaiyun Li, Dakui Wang, et al. 2021. FLOD: Oblivious Defender for Private Byzantine-Robust Federated Learning with Dishonest-Majority. In *ESORICS.*
[10] Peeyush Gupta, Yin Li, Sharad Mehrotra, Nisha Panwar, et al. 2019. Obscure: Information-Theoretic Oblivious and Verifiable Aggregation Queries. *PVLDB* (2019).
[11] Mohamad Mansouri, Melek Önen, Wafa Ben Jaballah, and Mauro Conti. 2023. SoK: Secure Aggregation Based on Cryptographic Schemes for Federated Learning. *PETS* (2023).
[12] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. PMLR.
[13] Julien Mirval, Luc Bouganim, and Iulian Sandu-Popa. 2021. Practical Fully-Decentralized Secure Aggregation for Personal Data Management Systems. In *SSDBM.*
[14] Julien Mirval, Luc Bouganim, and Iulian Sandu Popa. 2023. Federated Learning on Personal Data Management Systems: Decentralized and Reliable Secure Aggregation Protocols. In *Proceedings of the 35th International Conference on Scientific and Statistical Database Management.* 1–12.
[15] Amaury Bouchra Pilet, Davide Frey, and François Taïani. 2019. Robust Privacy-Preserving Gossip Averaging. In *SSS.*

Découverte de vérité confidentielle par calcul multi-parti

Angelo Saadeh
angelo.saadeh@telecom-paris.fr
LTCI, Télécom Paris, IP Paris &
CNRS@CREATE LTD
Palaiseau, France

Pierre Senellart
pierre@senellart.com
DI ENS, ENS, CNRS, Université PSL &
Inria & IUF & CNRS@CREATE LTD
& IPAL, CNRS
Paris, France

Stéphane Bressan
steph@nus.edu.sg
National University of Singapore &
CNRS@CREATE LTD & IPAL, CNRS
Singapore, Singapore

RÉSUMÉ

Un défi de la distribution de la découverte de connaissance et de la fouille de données est d'estimer la fiabilité des données provenant de sources autonomes, tout en protégeant la confidentialité de ces sources. Les algorithmes de découverte de vérité aident à corroborer les données de sources contradictoires. Pour chaque requête reçue, un algorithme de découverte de vérité prédit une valeur de vérité de la réponse, en mettant éventuellement à jour le score de confiance de chaque source. Peu de travaux, cependant, s'intéressent aux problèmes de confidentialité. Nous concevons et présentons un protocole de calcul multi-parti sécurisé basé sur du partage de secret, afin de réaliser des tests de pseudo-égalité qui sont utilisés dans des algorithmes de découverte de la vérité pour calculer des additions en fonction d'une condition. Le protocole garantit la confidentialité des données et des sources. Nous présentons également des variantes des algorithmes de découverte de la vérité qui sont rendues plus rapides quand on utilise du calcul multi-parti. Nous évaluons de manière empirique la performance du protocole proposé sur deux algorithmes de découverte de la vérité qui font partie de l'état de l'art, Cosine et 3-Estimates, et les comparant avec des algorithmes sans calcul multi-parti. Les résultats confirment que les algorithmes utilisant le calcul multi-parti sécurisé à base de partage de secret sont aussi précis que les versions standard, si on fait exception d'approximations numériques qui permettent de réduire la complexité du calcul.

VERSION COMPLÈTE

Ce travail est décrit dans un article en anglais qui a été présenté à la conférence DEXA 2023 [2] ; une version étendue est également disponible [1].

Remerciements. Ces recherches font partie du programme DesCartes et sont soutenues par la *National Research Foundation*, bureau du premier ministre, Singapour, au sein de son programme *Campus for Research Excellence and Technological Enterprise (CREATE).*

RÉFÉRENCES

[1] Angelo Saadeh, Pierre Senellart, and Stéphane Bressan. 2023. Confidential Truth Finding with Multi-Party Computation. In *Proc. DEXA*. Penang, Malaysia.
[2] Angelo Saadeh, Pierre Senellart, and Stéphane Bressan. 2023. Confidential Truth Finding with Multi-Party Computation (Extended Version). CoRR abs/2305.14727.

Appliance Detection Using Very Low-Frequency Smart Meter Time Series

Adrien Petralia
EDF - Université Paris Cité
Paris, France
adrien.petralia@gmail.com

Philippe Charpentier
EDF
Palaiseau, France
philippe.charpentier@edf.fr

Paul Boniol
Université Paris Cité
Paris, France
boniol.paul@gmail.com

Themis Palpanas
Université Paris Cité - IUF
Paris, France
themis@mi.parisdescartes.fr

KEYWORDS

Appliance Detection, Smart Meter Data, Time Series Classification

1 EXTENDED ABSTRACT

The energy sector is undergoing significant changes, primarily driven by the need for a more sustainable and secure energy supply. One way to better manage our consumption is to understand it better. In the last decade, electricity suppliers have installed millions of smart meters worldwide to improve their ability to manage the electrical grid [4, 10]. These meters record detailed time-stamped data on electricity consumption, allowing both individual customers and businesses to better understand and rationalize their consumption [3]. These data are also valuable for suppliers, as they can help them anticipate energy demand more accurately. Overall, the widespread adoption of smart meters plays a crucial role in transitioning toward a more sustainable and efficient energy system.

We note that it has become essential for electricity suppliers to know which electrical appliances their customers own. This knowledge allows suppliers to better segment their customer base [1], and therefore to propose personalized offers and services that increase the customer satisfaction and retention. Furthermore, they can help customers rationalize their electricity consumption, therefore contributing to the energy transition. One way to gather this information is by asking customers directly through a consumption questionnaire. However, this method can be a significant investment in terms of time and resources, which customers may not accept, and is also prone to errors. Therefore, electricity suppliers need to find more efficient and non-intrusive ways of gathering this information, such as using advanced data analytics techniques to detect the appliances directly through the collected smart meters data [5].

Appliance detection has become a significant area of research, with various techniques employed to detect the presence of devices [9, 12]. This problem is closely related to Non-Intrusive Load Monitoring (NILM), which aims to identify the power consumption, pattern, or on/off state activation of individual appliances using only the total consumption series [7]. While detecting an appliance can be seen as a step in NILM-based methods [2, 6, 8, 11], and diverse approaches have been proposed in the literature [2, 6, 8, 11], they differ from our objective. Indeed, these studies essentially focus on detecting *when* a specific appliance is "ON" rather than *if* a household owns a specific appliance, and the presence of a specific appliance is in several cases already known before applying

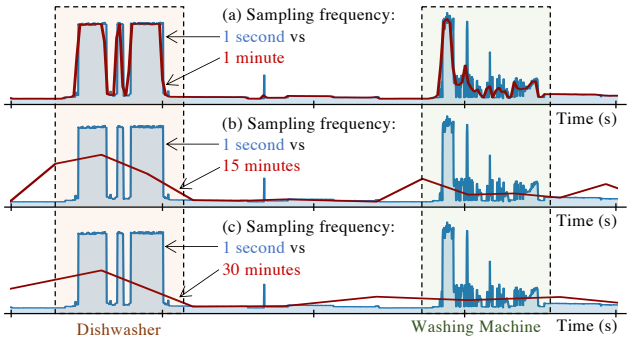


Figure 1: Comparisons of load curves containing a dishwasher and a washing machine at different sampling frequencies (1 second vs 1, 15, and 30min)

these approaches. Moreover, the majority of the NILM studies rely on data sampled at ≥1Hz, and consequently use signature-based methods [9, 12] that require either knowledge about how each appliance operates, or training on their individual power consumption. Nonetheless, most existing smart meter installations record consumption at a very low sampling frequency: once every 10 to 60 minutes (in some cases at an even lower frequency). This results in signals where the unique appliance pattern information has been smoothed-out, or lost. Figure 1 illustrates this loss of information. We observe that the dishwasher (shown on the left) and washing machine (shown on the right) signatures become increasingly hard to distinguish from one another as the sampling frequency drops. Therefore, it becomes infeasible to accurately detect appliances using signature-based methods for the sampling frequencies actually used in practice.

In this paper, we propose a benchmark of diverse state-of-the-art classification methods for the problem of appliance detection in very low-frequency electrical consumption time series. We conduct our experimental evaluation on five real smart meter datasets using different time series classifiers. We first focus on detecting appliances in very low-sampled smart meters data (30min level), as it is nowadays one of the standard sampling rates adopted by electricity suppliers. We then provide an in-depth analysis of the increasing detection quality using higher frequency smart meter readings: 15min, 10min, and 1 min. To our knowledge, this is the first study to perform an exhaustive comparison of 11 state-of-the-art methods on five diverse real datasets with 13 different types of appliances, for multiple sampling frequencies. The experimental evaluation demonstrates that current time series classifiers can accurately detect several appliances, even at the 30min resolution. Specifically, deep learning techniques are the most accurate and scalable when applied to large smart meter datasets. Moreover, we demonstrate that setting the smart meter reading frequency to 1min

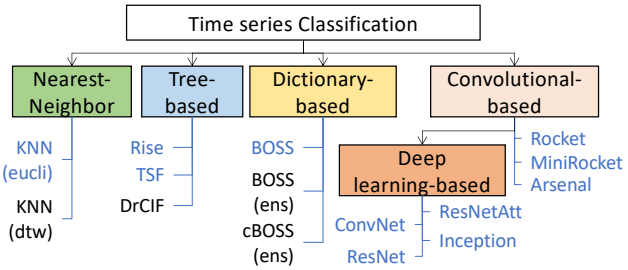


Figure 2: Taxonomy of classifier considered in our benchmark (in blue: classifier used in the experimental evaluation).

can greatly enhance appliance detection using time series classifiers. Our contributions are summarized as follows.

- We describe a framework for comparing the performance of different time series classification methods for the appliance detection problem and make this framework publicly available: https://github.com/adrienpetralia/ApplianceDetectionBenchmark
- We perform an extensive experimental evaluation using 5 diverse real datasets and 11 time series classifiers, including both traditional machine learning, as well as deep learning methods.
- We report the results of our comparison, which demonstrate that (i) current time series classifiers can only detect certain appliances at the 30min resolution; (ii) deep learning classifiers are the most accurate and scalable solution; and (iii) electricity suppliers should target a minimum smart meter reading frequency of 15min.
- The findings of this study can help electricity suppliers make informed decisions regarding the characteristics of future smart meter deployments. Moreover, these findings point to interesting (and still challenging) open research directions in the context of electricity consumption time series analysis and appliance detection in particular.

The full version of this paper is available at https://dl.acm.org/doi/10.1145/3575813.3595198.

2 EXPERIMENTAL EVALUATION

The different approaches proposed in the literature to solve the time series classification problem are shown in Figure 2. The objective is to compare the performance of these methods when applied to the appliance detection problem.

The overall results using 30min sampled data, shown in Figure 3, demonstrate that InceptionTime outperforms other classifiers when considering the average score and rank; InceptionTime is followed by ResNet, Arsenal, ConvNet, MiniRocket, and Rocket.

Figure 4 summarizes the average total running time (i.e., training and inference time together) for the 11 classifiers we studied. Taking into consideration the performance of the convolutional-based approaches (deep- and non-deep-learning approaches), as well as their running time, we observe that this type of classifier is the most suitable for appliance detection using 30min sampled smart meter data. InceptionTime reaches a slightly higher detection score, but at the cost of longer execution times. A balance between performance and efficiency is achieved by the ResNet and ConvNet classifiers.

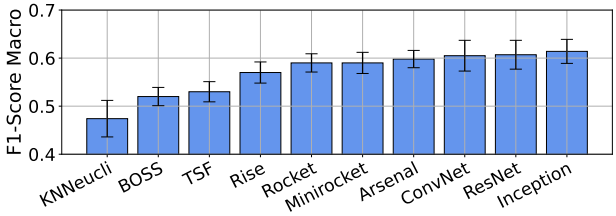


Figure 3: Average classifiers detection score through all the detection cases and all the datasets.

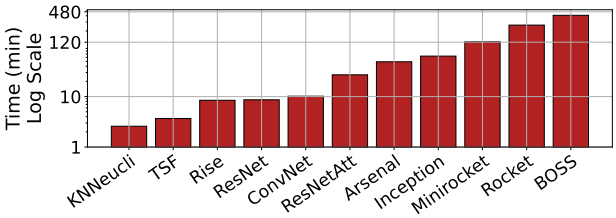


Figure 4: Average running time per run (training + inference time) for all classifiers (log scale y-axis).

REFERENCES

[1] Hunt Allcott. 2011. Social norms and energy conservation. *Journal of Public Economics* 95, 9 (2011), 1082–1095. https://doi.org/10.1016/j.jpubeco.2011.03.003 Special Issue: The Role of Firms in Tax Systems.

[2] Muzaffer Aslan and Ebra Nur Zurel. 2022. An efficient hybrid model for appliances classification based on time series features. *Energy and Buildings* 266 (2022), 112087. https://doi.org/10.1016/j.enbuild.2022.112087

[3] Gouri R. Barai, Sridhar Krishnan, and Bala Venkatesh. 2015. Smart metering and functionalities of smart meters in smart grid - a review. In *2015 IEEE Electrical Power and Energy Conference (EPEC)*. 138–145. https://doi.org/10.1109/EPEC.2015.7379940

[4] Stanislav Chren, Bruno Rossi, and Tomáš Pitner. 2016. Smart grids deployments within EU projects: The role of smart meters. In *2016 Smart Cities Symposium Prague (SCSP)*. 1–5. https://doi.org/10.1109/SCSP.2016.7501033

[5] G.W. Hart. 1992. Nonintrusive appliance load monitoring. *Proc. IEEE* 80, 12 (1992), 1870–1891. https://doi.org/10.1109/5.192069

[6] Matthias Kahl, Daniel Jorde, and Hans-Arno Jacobsen. 2022. Representation Learning for Appliance Recognition: A Comparison to Classical Machine Learning. https://doi.org/10.48550/ARXIV.2209.03759

[7] Maria Kaselimi, Eftychios Protopapadakis, Athanasios Voulodimos, Nikolaos Doulamis, and Anastasios Doulamis. 2022. Towards Trustworthy Energy Disaggregation: A Review of Challenges, Methods, and Perspectives for Non-Intrusive Load Monitoring. *Sensors* 22 (08 2022), 5872. https://doi.org/10.3390/s22155872

[8] Pauline Laviron, Xueqi Dai, Bérénice Huquet, and Themis Palpanas. 2021. Electricity Demand Activation Extraction: From Known to Unknown Signatures, Using Similarity Search. In *e-Energy '21: The Twelfth ACM International Conference on Future Energy Systems, Virtual Event, Torino, Italy, 28 June - 2 July, 2021*, Herman de Meer and Michela Meo (Eds.). ACM, 148–159. https://doi.org/10.1145/3447555.3464865

[9] Yu Liu, Congxiao Liu, Yiwen Shen, Xin Zhao, Shan Gao, and Xueliang Huang. 2021. Non-intrusive energy estimation using random forest based multi-label classification and integer linear programming. *Energy Reports* 7 (2021), 283–291. https://doi.org/10.1016/j.egyr.2021.08.045 2021 The 4th International Conference on Electrical Engineering and Green Energy.

[10] Megan Milam and G. Kumar Venayagamoorthy. 2014. Smart meter deployment: US initiatives. In *ISGT 2014*. 1–5. https://doi.org/10.1109/ISGT.2014.6816507

[11] Leitao Qu, Yaguang Kong, Meng Li, Wei Dong, Fan Zhang, and Hongbo Zou. 2023. A residual convolutional neural network with multi-block for appliance recognition in non-intrusive load identification. *Energy and Buildings* 281 (2023), 112749. https://doi.org/10.1016/j.enbuild.2022.112749

[12] Florian Rossier, Philippe Lang, and Jean Hennebert. 2017. Near Real-Time Appliance Recognition Using Low Frequency Monitoring and Active Learning Methods. *Energy Procedia* 122 (2017), 691–696. https://doi.org/10.1016/j.egypro.2017.07.371 CISBAT 2017 International ConferenceFuture Buildings & Districts – Energy Efficiency from Nano to Urban Scale.

ATEM: A Topic Evolution Model for the Detection of Emerging Topics in Scientific Archives

Hamed Rahimi, Hubert Naacke, Camelia Constantin, and Bernd Amann
LIP6, Sorbonne University
Paris, France

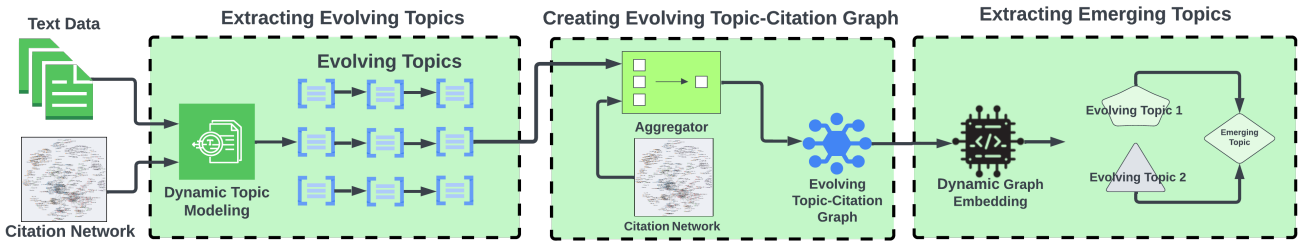


Figure 1: The Architecture of ATEM

ABSTRACT

This paper presents ATEM, a novel framework for studying topic evolution in scientific archives. ATEM employs dynamic topic modeling and dynamic graph embedding to explore the dynamics of content and citations within a scientific corpus. ATEM explores a new notion of citation context that uncovers emerging topics by analyzing the dynamics of citation links between evolving topics. Our experiments demonstrate that ATEM can efficiently detect emerging cross-disciplinary topics within the DBLP archive of over five million computer science articles.

KEYWORDS

Evolution Model, Topic Emergence, Science Evolution

1 MOTIVATION

Scientific research is a continuous process that generates new theories shaped by the collective efforts of scientists through research, experimentation, and analysis. Understanding the evolution of science has the capacity to revolutionize the research landscape, as it has significant implications for research funding and public policy decisions in academic and industrial settings. One of the most useful analyses of the evolution of science is the detection of topic emergence, which involves the identification of new areas of research and study within scientific disciplines. Emerging topics are ideas or issues that gain attention or become more prominent in a particular field or area of interest. The detection of emerging topics has far-reaching implications for society, providing a way to track the evolution of scientific fields and to shape future research and technological development.

There are many approaches to analyzing the evolution of science. Topic-based approaches can identify trends for specific terms or phrases but may not capture the broader context and relationships between scientific concepts, while citation-based approaches capture the relationships between scientific articles but are less effective at identifying trends for specific terms or phrases. These limitations highlight the need for a more holistic and versatile approach to provide a deeper understanding of topic evolution, while preserving the nuanced relationships between scientific topics. In this paper, we aim to discover emerging topics by proposing a framework, called ATEM, that discovers the evolution of science with different analyses. ATEM is driven by the recognition that citation links serve a dual purpose: they not only signify semantic connections between different topics, but also suggest the potential emergence of new topics within the cited interdisciplinary domains.

2 METHOD

As illustrated in Figure 1, ATEM is a general-purpose framework for modeling and analyzing the evolution of topics generated from scientific archives:

- Extracting Evolving Topics: The first layer extracts evolving topics from a corpus of documents using dynamic topic modeling. for describing the evolution of topics within a set of documents along different time periods.
- Creating Evolving Topic-Citation Graph: The second layer projects the structure of the document citation network into an evolving topic-citation graph. Our hypothesis is that citations between documents reflect additional interesting relationships between the topics discussed in those documents.
- Extracting Emerging Topics: The third layer applies a dynamic graph embedding method on the evolving topic-citation graph and defines emerging topics as the clusters of evolving topics with similar graph embeddings at a given time period. The assumption behind is that graph embeddings reflect the citation context of topics and new cross-disciplinary topics emerge from these clusters.

3 EXAMPLE

We generated emerging topics from the DBLP archive available at https://www.aminer.org/citation. Figure 3 shows the evolution of

Conference BDA'23, Montpellier, France,

the graph embedding neighborhood of topic T680C6 about "nearest neighbor classifier" (Figure 2) in 2013. We can see, for example, that topic T661C6 about "diabetic, meal and blood_glucose", appears in 2013 as a close embedding neighbor of the evolving topic T680C6. Table 1 shows the documents shared by T680C6 and T661C6. These documents are obtained by taking the intersection of the results of two search queries $R(T680C6)$ =['nearest neighbors', 'knn', 'nearest neighbor'] and $R(T661C6)$ =['glycemic', 'hypoglycemia', 'hyperglycemia'] over the archive ranked by the average search score. The result shows that most of the top relevant documents for the emerging topic (T680C6, T661C6) were published after its emergence period in 2013.

Figure 2: Evolving topic T680C6.

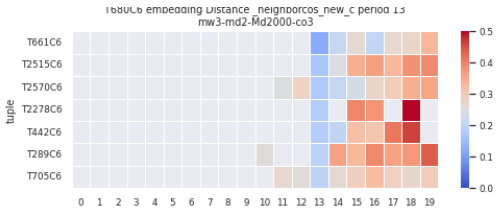


Figure 3: Embedding distance of topic T680C6 at period 2013

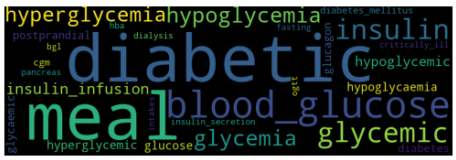


Figure 4: Evolving topic T661C6

Table 1: Emerging documents of topics (T680C6,T661C6)

Year	Title
2020.0	Performance evaluation of classification methods with PCA and PSO for diabetes.
2020.0	An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy
2020.0	Using Machine Learning to Predict the Future Development of Disease
2019.0	Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus.
2018.0	Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers.
2017.0	Automatic Diagnosis Metabolic Syndrome via a k- Nearest Neighbour Classifier.
2016.0	Predicting risk of suicide using resting state heart rate.
2015.0	Computer-aided diagnosis of diabetic subjects by heart rate variability signals using discrete wavelet transform method
2013.0	Automated detection of diabetes using higher order spectral features extracted from heart rate signals

4 VALIDATION

To validate our approach, the emerging topics generated by ATEM are compared to topic pairs freshly connected by citation links at a given time period (baseline). The emergence of embedding neighbors and citation neighbors is quantified by a new *emergence predictability* metrics comparing the number of documents published before and after the emergence period:

$$\mathcal{E}(t_e) : \frac{|D_{future}(t_e)| - |D_{past}(t_e)|}{|D(t_e)|} \quad (1)$$

where $D_{past}(t_e)$ are the documents published before the emergence period of t_e, and $D_{future}(t_e)$ are the documents published after the emergence period of t_e.

Figure 5 compares the predictability values for ATEM emerging topics (blue) and citation neighbors (orange). We observe that the emerging topics of ATEM have higher predictability compared to the baseline. Figure 6 shows the violin distribution of predictability values. As we can see, more than 50% of embedding neighbors define emerging topics with emergence value 0.5 (1.5/0.5 = 3 times more documents are published after emergence period than before) whereas this is only the case for less than 20% of citation neighbors.

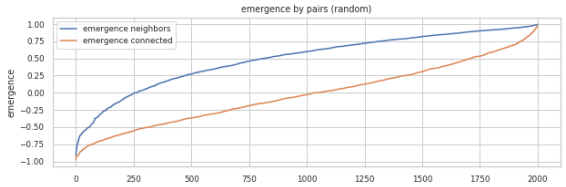


Figure 5: Average emergence values of ATEM (blue) vs co-citation analysis (orange)

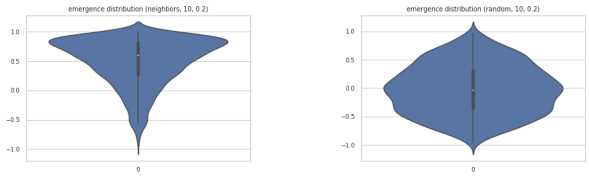


Figure 6: Emergence predictability distribution.

This article will be published in *The 12th International Conference on Complex Networks and their Applications* 28 - 30 November, 2023, Menton Riviera, France. See https://arxiv.org/abs/2306.02221 for a preprint with citations.

ANTM: An Aligned Neural Topic Model for Exploring Evolving Topics

Hamed Rahimi, Hubert Naacke, Camelia Constantin, and Bernd Amann
LIP6, Sorbonne University
Paris, France

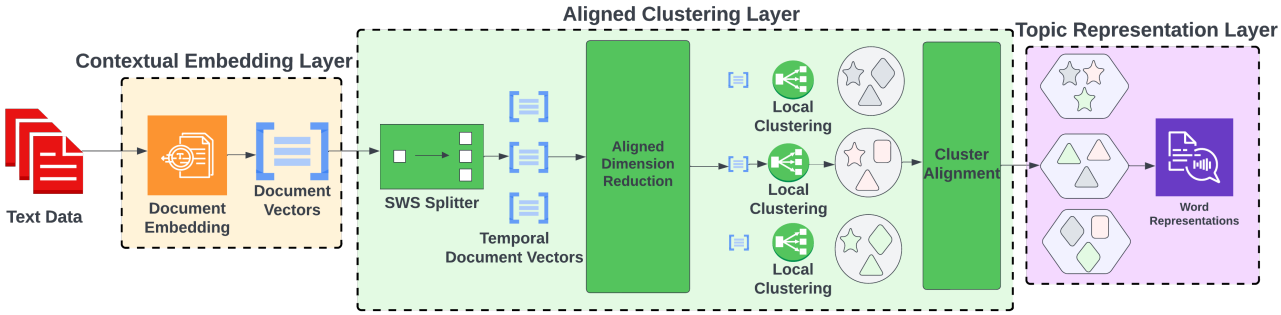


Figure 1: The Architecture of ATEM

ABSTRACT

This paper introduces ANTM, an algorithmic family of dynamic topic models that combines novel techniques for discovering evolving topics in large corpora. ANTM preserves the temporal continuity of evolving topics by extracting temporal features from documents (using advanced pre-trained large language models) and by employing an overlapping sliding window algorithm for sequential document clustering. This clustering method identifies different numbers of topics within each time frame and aligns semantically similar document clusters across time periods. This process captures emerging and fading topics and allows for a more diverse and interpretable representation of evolving topics. We evaluate ANTM against four other dynamic topic models on three datasets and conclude that it outperforms the state-of-the-art approaches in terms of interpretability and diversity. Furthermore, we demonstrate its effectiveness in processing large corpora, while improving the scalability and adaptability of dynamic topic models for different domains.

KEYWORDS

Dynamic Topic Modeling, Algorithmic Topic Models, Evolving Topics

1 INTRODUCTION

Topic modeling is a statistical technique used in natural language processing to discover abstract themes from a corpus of text documents. These models are widely used in exploratory data analysis for organizing, understanding, and summarizing large amounts of text data. Dynamic topic models are the temporal variants of topic models that update their estimates of the underlying topics as new documents are added to the corpus. These models analyse topic evolution and can be used to identify patterns in temporal archives. The application of these models includes discovering innovations in scientific archives and understanding trends in public opinion on particular issues.

2 MOTIVATION

Dynamic topic models are widely used to analyze topic evolution. However, they become computationally expensive when dealing with large archives covering extensive vocabularies of terms. Furthermore, they assume that topics evolve smoothly over time and are unable to capture abrupt semantic changes or topic drifts that may occur during paradigm shifts. BERTopic is a recent clustering-based topic model that addresses these issues. However, BERTopic has certain limitations. The primary concern relates to its method for building evolving topics, which relies on static document clusters with dynamic word representations. This limits BERTopic to maintaining the same number of subtopics in each time period, which leads to model instability if the document distribution within a dataset is not normalized. There are also cases where documents appear in the same global topics even though they belong to different subtopics within a given time period. These problems complicate the ability to track the dynamic progression of evolving topics, including shifts in the number of document clusters or cluster sizes as topics emerge and fade over time. It also reduces the explanatory and interpretative power of the model and fails to accurately represent evolving topics within a given time period.

3 METHOD

Compared to BERTopic, ANTM proposes to discover evolving topics through an overlapping sliding window algorithm for temporal document clustering. This method takes into account the changes in document content over time and produces a set of high-quality

topics for each time period. As shown in Figure 1, ANTM consists of three layers. The first layer uses advanced pre-trained LLMs to provide a time-aware vector representation for each document, corresponding to its content. The second layer splits the document vector representations into a set of temporal time frames and applies an overlapping sliding window for temporal document segmentation and clustering. Finally, the third layer is responsible for providing word representations for each set of aligned clusters over time.

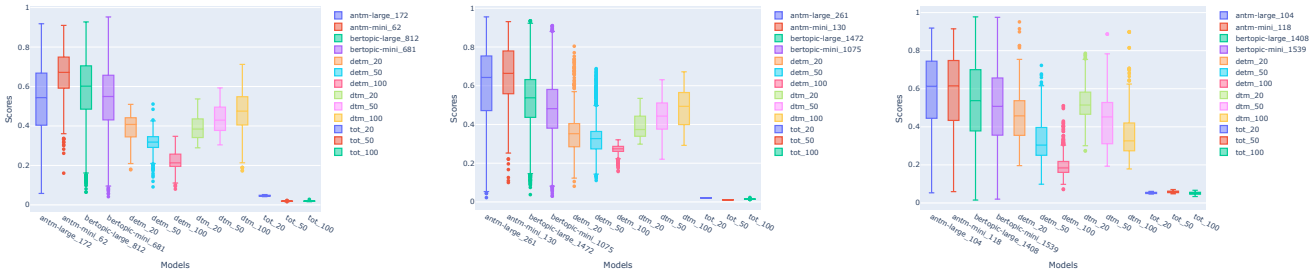
Figure 2: Evolving Topic about Ebola Outbreak.

Figure 2 illustrates the evolving topic about an Ebola outbreak, obtained through HDBSCAN alignment across four time periods. Slight adjustments can be observed in the representation of topics over time, while numerous topic words are reiterated (as indicated by grey dots) in the following time periods.

4 VALIDATION

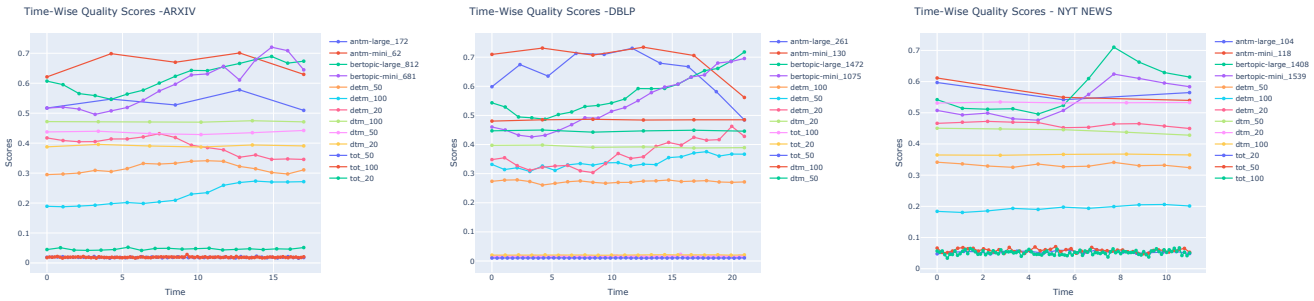
We conducted our experiments on three datasets and four dynamic topic models. The result demonstrate a significant improvement in coherence and diversity score compared to the baseline models. The effectiveness of ANTM can be evaluated from two angles. Firstly, we observe the quality of topics in terms of *coherence and diversity within each time frame* (period-wise analysis as shown in Figure 4). The second perspective examines the *quality of word representations in every evolving topic* (topic-wise analysis as shown in Figure 3) and compares the ability of dynamic topic models to represent the evolution of each dynamic topic over time. ANTM achieves the highest average Topic Quality (TC×TD) scores among the different baseline variants in all three datasets. However ANTM consumes more runtime compared to BERTopic due to the increased number of topic clustering steps. Yet, it remains significantly faster than D-ETM, DTM, and TOT.

An extended preprint of this work is available here: https://arxiv.org/abs/2302.01501



(a) ArXiv **(b) DBLP** **(c) NYT News**

Figure 3: Topic-wise Quality Comparison



(a) ArXiv **(b) DBLP** **(c) NYT News**

Figure 4: Period-wise Quality Comparison

Conjunctive Queries With Self-Joins, Towards a Fine-Grained Enumeration Complexity Analysis

Nofar Carmeli[1] and Luc Segoufin[2]

[1] DI ENS, ENS, Université PSL, CNRS, Inria Paris, France
[2] INRIA, ENS Paris, PSL Paris, France

Abstract. Even though query evaluation is a fundamental task in databases, known classifications of conjunctive queries by their fine-grained complexity only apply to queries without self-joins. We study how self-joins affect enumeration complexity, with the aim of building upon the known results to achieve general classifications. We do this by examining the extension of two known dichotomies: one with respect to linear delay, and one with respect to constant delay after linear preprocessing. As this turns out to be an intricate investigation, this paper is structured as an example-driven discussion that initiates this analysis. We show enumeration algorithms that rely on self-joins to efficiently evaluate queries that otherwise (i.e., if the relation names were replaced to eliminate self-joins) cannot be answered with the same guarantees. Due to these additional tractable cases, the hardness proofs are more complex than the self-join-free case. We show how to harness a known tagging technique to prove hardness of queries with self-joins. Our study offers sufficient conditions and necessary conditions for tractability and settles the cases of queries of low arity and queries with cyclic cores. Nevertheless, many cases remain open.

Efficient Enumeration of Recursive Plans in Transformation-based Query Optimizers

Amela Fejza
amela.fejza@inria.fr
Tyrex team, Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble
France

Pierre Genevès
pierre.geneves@inria.fr
Tyrex team, Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble
France

Nabil Layaïda
nabil.layaida@inria.fr
Tyrex team, Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble
France

ABSTRACT

Query optimizers built on the transformation-based Volcano/Cascades framework are used in many database systems. Transformations proposed earlier on the logical query dag (LQDAG) data structure, which is key in such a framework, focus only on recursion-free queries. In this paper, we propose the recursive logical query dag (RLQDAG) which extends the LQDAG with the ability to capture and transform recursive queries, leveraging recent developments in recursive relational algebra. Specifically, this extension includes: (i) the ability of capturing and transforming sets of recursive relational terms thanks to (ii) annotated equivalence nodes used for guiding transformations that are more complex in the presence of recursion; and (iii) RLQDAG rewrite rules that transform sets of subterms in a grouped manner, instead of transforming individual terms in a sequential manner; and that (iv) incrementally update the necessary annotations. Core concepts of the RLQDAG are formalized using a syntax and formal semantics with a particular focus on subterm sharing and recursion. The result is a clean generalization of the LQDAG transformation-based approach, enabling more efficient explorations of plan spaces for recursive queries. An implementation of the proposed approach shows significant performance gains compared to the state-of-the-art.

Query Rewriting with Disjunctive Existential Rules and Mappings

Michel Leclère, Marie-Laure Mugnier, and Guillaume Pérution-Kihli

LIRMM, Inria, University of Montpellier, CNRS, France

Abstract. We consider the issue of answering unions of conjunctive queries (UCQs) with disjunctive existential rules and mappings. While this issue has already been well studied from a chase perspective, query rewriting within UCQs has hardly been addressed yet. We first propose a sound and complete query rewriting operator, which has the advantage of establishing a tight relationship between a chase step and a rewriting step. The associated breadth-first query rewriting algorithm outputs a minimal UCQ-rewriting when one exists. Second, we show that for any "truly disjunctive" nonrecursive rule, there exists a conjunctive query that has no UCQ-rewriting. It follows that the notion of finite unification sets (fus), which denotes sets of existential rules such that any UCQ admits a UCQ-rewriting, seems to have little relevance in this setting. Finally, turning our attention to mappings, we show that the problem of determining whether a UCQ admits a UCQ-rewriting through a disjunctive mapping is undecidable. We conclude with a number of open problems.

5 Résumés des articles courts

Efficient Computation of General Modules for ALC Ontologies

Hui Yang
yang@lri.fr
LISN, Univ. Paris-Sud, CNRS, Université Paris-Saclay
Orsay, France

Patrick Koopmann
p.k.koopmann@vu.nl
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

Yue Ma
ma@lri.fr
LISN, Univ. Paris-Sud, CNRS, Université Paris-Saclay
Orsay, France

Nicole Bidoit
nicole.bidoit@lri.fr
LISN, Univ. Paris-Sud, CNRS, Université Paris-Saclay
Orsay, France

We present a method for extracting general modules for ontologies formulated in the description logic $\mathcal{ALC}$. A module for an ontology is an ideally substantially smaller ontology that preserves all entailments for a user-specified set of terms. As such, it has applications such as ontology reuse and ontology analysis. Different from classical modules, general modules may use axioms not explicitly present in the input ontology, which allows for additional conciseness. However, they still need to be entailed by the original ontology, and ideally should be substantially smaller. So far, general modules have only been investigated for lightweight description logics [1, 8]. Our main contributions are: 1) the first method dedicated to general modules in $\mathcal{ALC}$, 2) a formal analysis of some properties of the general modules we compute, 3) new methods for module extraction and uniform interpolation that significantly improve the state-of-the-art, 4) an evaluation on real-world ontologies indicating the efficiency of our technique. This work has been accepted by IJCAI 2023. For detailed results and proofs, please refer to the extended version of the paper [11].

The main steps of our approach are shown in Figure 1. Similar to [6], our method essentially works by performing uniform interpolation on a normalized version of the input ontology. During normalization, so-called *definer names* are introduced, which are eliminated in the final step. However, different from [6], we put fewer constraints on the normal form and do not allow the introduction of definers after normalization, which changes the mechanism of uniform interpolation. As a result, our definer elimination step may reintroduce names eliminated during uniform interpolation, which is not a problem for the computation of general modules. In contrast, eliminating definers as done in [6] can cause an exponential blowup, and introduce concepts with the non-standard *greatest fixpoint* constructor [2].

Ontology Normalization As the first step, we normalize any input ontology $\mathcal{O}$ by standard transformations as in [5]. An ontology $\mathcal{O}$ is in *normal form* if every axiom is of the following form, where A is a concept name:

$$\top \sqsubseteq L_1 \sqcup \ldots \sqcup L_n \qquad L_i ::= A \mid \neg A \mid \mathsf{Q}r.A, \;\; \mathsf{Q} \in \{\forall, \exists\}.$$

For simplicity, we omit the "$\top \sqsubseteq$" on the left-hand side of normalized axioms.

Role Forgetting Next, we apply *role forgetting* to eliminate concept names outside a given *signature* Σ (i.e., a set of concept and role names that one interested in). Existing methods to compute role forgetting either rely on an external reasoner [6, 12] or use the *universal role* ∇ [7, 13]. The former approach can be expensive, while the latter produces axioms outside of $\mathcal{ALC}$. The normalization allows us to implement a more efficient solution within $\mathcal{ALC}$, which relies on an integrated reasoning procedure and an additional transformation step that produces so-called *role isolated ontologies* $RI_\Sigma(\mathcal{O})$ for each input ontology $\mathcal{O}$ and signature Σ. For role isolated ontologies, role forgetting is straighforward by the following result.

Theorem 1. *Let* $\mathtt{rolE}_\Sigma(\mathcal{O})$ *be the ontology obtained as follows: (1) apply the* r-Rule *in Figure 2 exhaustively for each* $r \in sig_{\mathsf{R}}(\mathcal{O}) \setminus \Sigma$*, and then (2) remove all axioms containing some* $r \in sig_{\mathsf{R}}(\mathcal{O}) \setminus \Sigma$*. If* $\mathcal{O}$ *is role isolated for* Σ*, then* $\mathtt{rolE}_\Sigma(\mathcal{O})$ *is a role forgetting for* $\mathcal{O}$ *and* Σ*.*

Concept Forgetting Inspired by [13, Theorem 1], we define a concept forgetting operator $\mathtt{conE}_\Sigma$ using the *A-Rule* in Figure 2 in a similar way as we defined $\mathtt{rolE}_\Sigma$. Applying the aforementioned procedure yields $\mathtt{conE}_\Sigma(\mathtt{rolE}_\Sigma(RI_\Sigma(\mathcal{O})))$, which contains only names in the signature Σ and definers.

Constructing the General Module Now, to obtain our general modules $gm_\Sigma(\mathcal{O})$ for $\mathcal{O}$ and Σ, we have to eliminate the definers D from $\mathtt{conE}_\Sigma(\mathtt{rolE}_\Sigma(RI_\Sigma(\mathcal{O})))$. This is done by replacing each definer D by C_D, which is the concept replaced by D in the normalization step. To improve the results, we eliminate some definers before substituting them. In particular, inspired by [10], we apply some optimization operations on $\mathtt{conE}_\Sigma(\mathtt{rolE}_\Sigma(RI_\Sigma(\mathcal{O})))$. The produced optimized general modules are denoted by $gm^*_\Sigma(\mathcal{O})$.

Deductive Modules and Uniform Interpolants Some applications require the constraints of classical modules (being a subset of the original ontology) or of uniform interpolants (using only names from the signature). Those can be computed using our method as well. We can compute a deductive module $dm_\Sigma(\mathcal{O})$ for $\mathcal{O}$ and Σ by tracing back the inferences performed when computing the general module $gm^*_\Sigma(\mathcal{O})$. If instead of substituting definers D by C_D, we eliminate them using existing uniform interpolation tools such as Lethe or Fame [6, 14], we can compute a uniform interpolant for the input.

Evaluation We use 222 ontologies $\mathcal{ALC}$ ontologies that are generated from the OWL Reasoner Evaluation (ORE) 2015 classification track [9] by removing axioms not expressible in $\mathcal{ALC}$. For

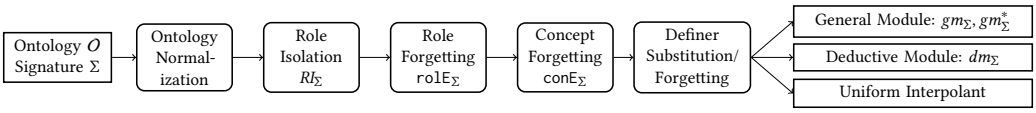


Figure 1: Overview of our unified method for computing general modules, deductive modules and uniform interpolants.

$$\underline{A\text{-}Rule}: \quad \frac{C_1 \sqcup A_1 \qquad \neg A_1 \sqcup C_2}{C_1 \sqcup C_2}$$

$$\underline{r\text{-}Rule}: \quad \frac{C_1 \sqcup \exists r.D_1,\ \bigcup_{j=2}^{n}\{C_j \sqcup \forall r.D_j\},\ K_D}{C_1 \sqcup \ldots \sqcup C_n},$$

where $K_D = \neg D_1 \sqcup \ldots \sqcup \neg D_n$ or $\neg D_2 \sqcup \ldots \sqcup \neg D_n$.

Figure 2: Inference rules for computing $\mathcal{D}_\Sigma(\mathcal{O})$

Table 1: Comparison of different methods (max./avg./med.).

Methods	Success rate	Resulting ontology length	Time cost
minM	84.34%	2,355 / 392.59 / 264	595.88 / 51.82 / 8.86
$\top\bot^*$-module	100%	4,008 / 510.77 / 364	5.94 / 1.03 / 0.90
FAME	91.25%	9,446,325 / 6,661.01 / 271	526.28 / 3.20 / 1.17
LETHE	85.27%	131,886 / 609.30 / 196	598.20 / 49.21 / 13.57
GEMO gm	97.34%	179,999 / 2,335.05 / 195	17.50 / 2.44 / 1.63
GEMO gm*		21,891 / 466.15 / 166	
GEMO dm		2,789 / 366.36 / 249	
gmLethe	96.17%	21,891 / 364.10 /162	513.15 / 3.08 / 1.68

each ontology, we randomly generated 50 signatures consisting of 100 concept and role names as in [7]. We implemented a prototype called GEMO in Python 3.7.4. For each request $(\mathcal{O}, \Sigma)$, GEMO produced three different (general) modules $gm_\Sigma(\mathcal{O})$, $gm^*_\Sigma(\mathcal{O})$ and $dm_\Sigma(\mathcal{O})$, respectively denoted by gm, gm*, and dm. gmLethe denotes the uniform interpolation method described above, where we used GEMO for computing gm* and then LETHE for definer forgetting.

To show that our general modules can serve as a better alternative for ontology reuse and analysis, we compared them with the state-of-the-art tools implementing module extraction and uniform interpolation for $\mathcal{ALC}$: **(i)** $\top\bot^*$*-modules* [3] as implemented in the OWL API [4]; **(ii)** minM [7] that computes *minimal deductive modules* under $\mathcal{ALCH}^\nabla$-semantics; **(iii)** LETHE 0.6[1][6] and FAME 1.0[2] [14] that compute uniform interpolants. Some of results are shown below.

- *Success rate:* We say a method *succeeds* on a request if it outputs the expected results within 600s. From Table 1 we can see that, after the $\top\bot^*$-modules, our method GEMO had the highest success rate.
- *Resulting ontology length and run time:* Because some of the methods can change the shape of axioms, the number of axioms is not a good metric for understanding the quality of general modules. We thus chose to use ontology length, which counts the number of concept and role name appearances in the ontology, for our evaluation. Table 1 shows the length and run time for the requests on which all methods were successful (78.45% of all requests). We observe that (i) dm and gmLethe have the best overall performance: their results had a substantially smaller average length and were computed significantly faster than others; (ii) Comparing gm and gm* regarding length lets us conclude that our optimization is effective.

[1] https://lat.inf.tu-dresden.de/~koopmann/LETHE/
[2] http://www.cs.man.ac.uk/~schmidt/sf-fame/

REFERENCES

[1] Ghadah Alghamdi, Renate A. Schmidt, Warren Del-Pinto, and Yongsheng Gao. 2021. Upwardly Abstracted Definition-Based Subontologies. In *K-CAP '21: Knowledge Capture Conference*, Anna Lisa Gentile and Rafael Gonçalves (Eds.). ACM, 209–216. https://doi.org/10.1145/3460210.3493564

[2] Diego Calvanese and Giuseppe De Giacomo. 2003. Expressive description logics. In *The description logic handbook: theory, implementation, and applications*. 178–218.

[3] B Cuenca Grau, Ian Horrocks, Yevgeny Kazakov, and Ulrike Sattler. 2008. Modular reuse of ontologies: Theory and practice. *Journal of Artificial Intelligence Research* 31 (2008), 273–318.

[4] Matthew Horridge and Sean Bechhofer. 2011. The OWL API: A Java API for OWL ontologies. *Semantic Web* 2, 1 (2011), 11–21. https://doi.org/10.3233/SW-2011-0025

[5] Patrick Koopmann. 2015. *Practical uniform interpolation for expressive description logics.* Ph.D. Dissertation. University of Manchester, UK. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.674705

[6] Patrick Koopmann. 2020. LETHE: Forgetting and uniform interpolation for expressive description logics. *KI—Künstliche Intelligenz* 34, 3 (2020), 381–387.

[7] Patrick Koopmann and Jieying Chen. 2020. Deductive Module Extraction for Expressive Description Logics. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Christian Bessiere (Ed.). 1636–1643. https://doi.org/10.24963/ijcai.2020/227

[8] Nadeschda Nikitina and Birte Glimm. 2012. Hitting the Sweetspot: Economic Rewriting of Knowledge Bases. In *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 7649)*, Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, Jérôme Euzenat, Manfred Hauswirth, Josiane Xavier Parreira, Jim Hendler, Guus Schreiber, Abraham Bernstein, and Eva Blomqvist (Eds.). Springer, 394–409. https://doi.org/10.1007/978-3-642-35176-1_25

[9] Bijan Parsia, Nicolas Matentzoglu, Rafael S. Gonçalves, Birte Glimm, and Andreas Steigmiller. 2017. The OWL Reasoner Evaluation (ORE) 2015 Competition Report. *J. Autom. Reason.* 59, 4 (2017), 455–482.

[10] Mostafa Sakr and Renate A. Schmidt. 2022. Fine-Grained Forgetting for the Description Logic $\mathcal{ALC}$. In *Proceedings of the 35th International Workshop on Description Logics (DL 2022) (CEUR Workshop Proceedings, Vol. 3263)*, Ofer Arieli, Martin Homola, Jean Christoph Jung, and Marie-Laure Mugnier (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-3263/paper-17.pdf

[11] Hui Yang, Patrick Koopmann, Yue Ma, and Nicole Bidoit. 2023. Efficient Computation of General Modules for $\mathcal{ALC}$ Ontologies (Extended Version). arXiv:2305.09503 [cs.AI] https://arxiv.org/abs/2305.09503.

[12] Yizheng Zhao, Ghadah Alghamdi, Renate A. Schmidt, Hao Feng, Giorgos Stoilos, Damir Juric, and Mohammad Khodadadi. 2019. Tracking Logical Difference in Large-Scale Ontologies: A Forgetting-Based Approach. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*. AAAI Press, 3116–3124.

[13] Yizheng Zhao and Renate A. Schmidt. 2017. Role Forgetting for $\mathcal{ALCOQH}(\nabla)$-Ontologies Using an Ackermann-Based Approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, Carles Sierra (Ed.). ijcai.org, 1354–1361.

[14] Yizheng Zhao and Renate A Schmidt. 2018. FAME: an automated tool for semantic forgetting in expressive description logics. In *International Joint Conference on Automated Reasoning*. Springer, 19–27.

Identification de données pertinentes dans des sources RDF

Zoé Chevallier
Université de Versailles Saint-Quentin-en-Yvelines
Versailles, France
zoe.chevallier@uvsq.fr

Zoubida Kedad
Université de Versailles Saint-Quentin-en-Yvelines
Versailles, France
zoubida.kedad@uvsq.fr

Béatrice Finance
Université de Versailles Saint-Quentin-en-Yvelines
Versailles, France
beatrice.finance@uvsq.fr

Frédéric Chaillan
Grand Paris Sud
Lieusaint, France
f.chaillan@grandparissud.fr

RÉSUMÉ

Le nombre croissant de sources de données RDF(S)/OWL qui sont publiées sur le Web représente une quantité sans précédent d'information disponible pour diverses applications. Mais pour en exploiter le potentiel, les sous-ensembles pertinents de ces sources pour un besoin spécifique doivent être identifiés, ce qui est complexe car le schéma décrivant ces sources n'est pas nécessairement fourni. Dans cet article, nous proposons une approche permettant d'identifier des instances candidates pour un schéma cible dans une source de données RDF dont le schéma est incomplet ou absent. Notre approche repose sur un algorithme d'apprentissage semi-supervisé qui compare de façon itérative les entités contenues dans les sources aux classes du schéma cible ainsi qu'aux entités candidates déjà identifiées pour ces classes.

CCS CONCEPTS

• **Information systems → Data extraction and integration**.

MOTS CLÉS

Extraction de données web, apprentissage semi-supervisé, données RDF

1 INTRODUCTION

Un très grand nombre de sources de données dans des domaines très variés sont publiées sur le Web, décrites dans les langages proposés par le W3C (World Wide Web Consortium), comme le langage RDF [5]. Une source de données RDF contient à la fois des données et le schéma qui les décrit. Ce shéma peut être incomplet ou même absent, et il ne représente pas une contrainte sur les données. Une ressource déclarée comme une instance d'une classe peut être décrite par un ensemble de propriétés différent de celui déclaré dans le schéma pour cette classe. Une ressource ne possède pas toujours de déclaration de type.

Si nous considèrons que les besoins d'une application spécifique sont exprimés à l'aide d'un schéma cible, défini en RDFS/OWL, l'identification des données pertinentes dans une source de données RDF pour laquelle le schéma est absent ou partiel revient à explorer cette source pour identifier quelles ressources pourraient être des instances candidates pour le schéma cible. Ce processus d'instanciation est complexe et consommateur de temps. Notre problème peut être formulé comme suit : étant donné une classe C définie dans le schéma cible, et S une source de données RDF, comment identifier dans S l'ensemble d'entités $I_C = \{e_1, e_2, ..., e_n\}$ tel que chaque entité e_i dans I_C soit une instance candidate pour C.

L'objectif du travail présenté ici consiste à peupler de façon automatique un schéma cible à partir de sources de données décrites en RDF. Nous proposons pour cela une approche fondée sur un algorithme d'apprentissage semi-supervisé qui identifie des entités candidates pour les classes du schéma cible de façon itérative. A chaque itération, les entités sources sont comparées aux descriptions des classes du schéma cible, mais également aux instances candidates déjà identifiées pour ces classes. Notre approche traite d'abord les entités sources typées, puis les entités sources non typées. Les entités représentant le même objet réel sont ensuite fusionnées pour ne pas avoir de doublons dans l'ensemble obtenu. Dans ce qui suit, nous présentons les différentes étapes de notre approche.

2 TRAITEMENT DES ENTITÉS SOURCES TYPÉES

Considérons une source de données S décrite en RDF, et un schéma cible T, décrit en RDFS/OWL, et notons $Match_{S,T}$ l'ensemble des correspondances entre T et le schéma de S. Pour chaque classe C_T du schéma cible, nous recherchons d'abord l'ensemble des instances candidates parmi les entités typées de la source S.

Une entité e de type C_S dans S est une instance candidate pour la classe cible C_T si $Eq_C(C_S, C_T) \in Match_{S,T}$, en d'autres termes, s'il existe une assertion d'équivalence entre C_S, le type de e, et C_T.

Dans notre exemple (cf. figure 1), le type déclaré pour l'entité source e_1 est la classe $s{:}Personne$. Comme nous avons $Match_{S,T} = \{Eq_C(\text{s:Personne, t:Personne})\}$, alors l'entité e_1 est une instance candidate pour $t{:}Personne$, $I_{t:Personne} = \{e1\}$.

3 TRAITEMENT DES ENTITÉS SOURCES NON TYPÉES

Pour déterminer si une entité non typée e dans une source est une instance candidate pour une classe C_T dans le schéma cible, e est comparée à la classe C_T, mais également aux instances candidates déjà identifiées pour C_T. Pour chaque nouvelle instance candidate identifiée pour C_T, les entités non typées de la source sont parcourues pour identifier de nouvelles instances candidates.

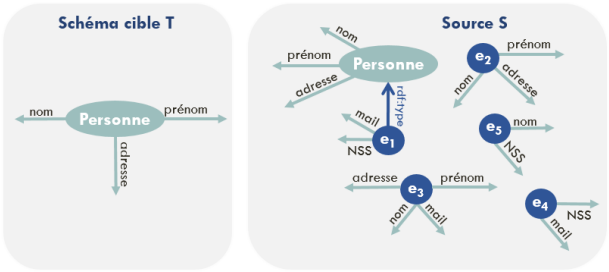


Figure 1 : Exemple de schéma cible et de source RDF

En effet, les nouvelles instances candidates peuvent introduire des propriétés et changer la fréquence des propriétés ce qui fait varier la mesure de similarité. Nous proposons d'adapter un algorithme d'apprentissage semi-supervisé (Self Training [6]) pour réaliser cette exploration itérative. Pour cela, nous utilisons deux métriques de similarité. La première évalue la similarité entre une entité et la description de la classe cible C_T. Elle est définir par l'indice de Jaccard [2] et compare l'ensemble de propriétés de la classe déclarées dans le schéma cible $Prop(C)$ à l'ensemble de propriétés caractérisant l'entité $Prop(e)$. La seconde métrique évalue la similarité entre une entité e et l'ensemble des instances candidates de C ; elle est fondée sur l'indice de Jaccard, mais intègre les fréquences des propriétés dans l'ensemble des instances candidates.

Dans notre exemple, considérons l'entité e_4 (cf. figure 1), qui a des propriétés différentes de celles de la classe *t:Personne*. Cependant, e_1 est une instance candidate de *t:Personne*, et e_4 à les mêmes propriétés que e_1. Par conséquent, e_4 est une instance candidate pour *t:Personne*.

Pour identifier les instances candidates parmi les entités sources non typées, nous adaptons un algorithme d'apprentissage semi-supervisé (Self Training [6]), qui prend en entrée, pour chaque classe cible C un ensemble d'entités candidates, noté I_C et un ensemble d'entités non typées. Parmi les entités non typées, les instances candidates sont identifiées à l'aide d'une probabilité d'appartenance d'une entité e à une classe C fondée sur les métriques de similarité schéma et instances. Si cette probabilité est supérieure à un seuil τ, alors e est une instance candidate pour C.

A chaque ajout dans l'ensemble d'entités candidates I_C, la fréquence des propriétés dans l'ensemble des instances candidates est modifié, et donc la valeur de la similarité instance change. Il est donc possible d'identifier de nouvelles instances candidates. A la fin d'une itération, si de nouvelles instances candidates ont été trouvées, une nouvelle itération est initiée. Le processus prend fin lorsqu'aucune nouvelle entité candidate n'est découverte.

4 FUSION DES ENTITÉS

Il est possible que deux entités e_i et e_j dans deux sources de données RDF distinctes fassent référence à un même objet réel. C'est le cas par exemple si une propriété *owl : sameAs* est définir pour e_i et a pour objet e_j. Si ces deux entités ont été identifiées comme des instances candidates pour la même classe du schéma cible, et pour ne pas introduire de doublons dans l'ensemble des instances, les deux entités sont fusionnées en une seule entité dont l'ensemble de propriété est obtenu en fusionnant les propriétés de e_i et e_j et en conservant la provenance des valeurs des propriétés.

5 TRAVAUX CONNEXES

L'identification d'informations pertinentes a fait l'objet de travaux de recherche comme les approches présentées dans [1, 3], qui proposent d'identifier les informations pertinentes pour enrichir une base cible en y ajoutant des tuples ou des propriétés, mais cette approche traite des données structurées dont la structure est connue. Une autre famille d'approches en lien avec notre travail sont les approches d'enrichissement de schéma, comme celle proposée dans [4], qui traitent l'inférence de type pour les entités non typées d'une source de données RDF. L'approche utilise une méthode statistique qui requiert un ensemble de définitions de types ainsi que les instances associées. Elle ne s'applique donc pas au cas où les types définis ne possèdent pas d'instances.

6 CONCLUSION

Nous avons proposé une approche permettant d'identifier des instances candidates des classes d'un schéma cible à partir d'une source dont le schéma est partiel ou absent. Pour les entités typées de la source, nous utilisons les correspondances existantes entre le schéma de la source et le schéma cible. Pour les entités non typées de la source, nous adaptons un algorithme d'apprentissage semi-supervisé permettant de comparer de manière itérative les entités aux classes du schéma cible. Lors de l'instanciation du schéma cible, nous proposons de fusionner les entités qui correspondent au même objet réel pour ne pas introduire de doublons.

Dans nos travaux futurs, nous étendrons nos algorithmes pour instancier les liens entre les classes cibles. Nous nous intéresserons également au passage à l'échelle de nos algorithmes pour traiter des sources de données de très grande taille.

REMERCIEMENTS

Ces travaux sont partiellement financés par l'action "*IA pour habiter le futur*" du projet Territoires d'Innovation de Grande Ambition (TIGA) de la région Île de France.

RÉFÉRENCES

[1] Alex Bogatu, Alvaro A. A. Fernandes, Norman W. Paton, and Nikolaos Konstantinou. 2020. Dataset Discovery in Data Lakes. In *36th IEEE International Conference on Data Engineering, ICDE 2020*. IEEE, 709–720.

[2] Paul Jaccard. 1901. Distribution de La Flore Alpine Dans Le Bassin Des Dranses et Dans Quelques Régions Voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles* 37 (1901), 241–272.

[3] Christos Koutras, George Siachamis, Andra Ionescu, Kyriakos Psarakis, Jerry Brons, Marios Fragkoulis, Christoph Lofi, Angela Bonifati, and Asterios Katsifodimos. 2021. Valentine : Evaluating Matching Techniques for Dataset Discovery. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)* (2021-04). IEEE, Chania, Greece, 468–479.

[4] Heiko Paulheim and Christian Bizer. 2013. Type Inference on Noisy RDF Data. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I* (2013) *(Lecture Notes in Computer Science, Vol. 8218)*. Springer, Sydney, NSW, Australia, 510–525.

[5] W3C. 1998. Resource Description Framework (RDF) Schemas. https://www.w3.org/TR/1998/WD-rdf-schema-19980409/

[6] David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics* -. Association for Computational Linguistics, Cambridge, Massachusetts, 189–196.

Overview and Perspectives for Optimistic JSON Schema Witness Generation

Lyes Attouche[1], Mohamed-Amine Baazizi[2], and Mimoun Malki[1]

[1] Université Paris-Dauphine, PSL Research University
[2] Sorbonne Université, LIP6 UMR 7606

Abstract. JSON Schema is an expressive schema language for describing JSON documents which combines structural and Boolean operators, and features negation and recursion. Satisfiability of JSON Schema is an important decision problem that can be solved by generating a witness when the schema admits one. However, the theoretical complexity of this problem is prohibitive [9] due to the combination of structural constructs, and logical operators, including negation. Optimistic witness generation is an alternative and attractive solution aimed at being fast while sacrificing completeness. Optimistic generation is an approach characterized by the fact that witnesses are first generated for each fragment of the schema hoping that their trivial combination, according to schema indications, form a global witness but does guarantee correctness, in the general case. To better understand the theoretical limitation of this approach, we build an incomplete yet correct witness generation solution and study its limitations on real-life schemas. We compare our solution with existing open-source solutions and identify potential improvements.

An End-to-End Machine Learning Framework for District Heating Networks Simulation

Taha Boussaid
taha.boussaid@insa-lyon.fr
Univ. Lyon, INSA Lyon,
CNRS, LIRIS, UMR 5205 & CETHIL, UMR 5008
Villeurbanne, FRANCE

Vasile-Marian Scuturici
Univ. Lyon, INSA Lyon,
CNRS, LIRIS, UMR 5205
Villeurbanne, FRANCE

François Rousset
Univ. Lyon, INSA Lyon,
CNRS, CETHIL, UMR 5008
Villeurbanne, FRANCE

Marc Clausse
Univ. Lyon, INSA Lyon,
CNRS, CETHIL, UMR 5008
Villeurbanne, FRANCE

ABSTRACT

Faced with environmental challenges, district heating networks (DHN) have been identified as a viable solution to decarbonize the heating sector. However, they raise various challenges regarding the optimization of their control given their size and the operational constraints of the energy systems involved. As a result, the numerical simulation of these networks is computationally heavy, which hinders near-instantaneous optimal control. In this work, we present the first building block of an optimization framework for the control of district heat networks using a surrogate model based on geometric deep learning. More precisely we trained specific architectures of Graph Neural Networks to emulate a thermo-hydraulic simulator of district heating network. This statistical inference method allows us to drastically reduce simulation time by 1 to 4 orders of magnitude.

1 INTRODUCTION

A detailed review on control strategies for DHN can be found in [2]. It shows that all strategies make use of several simulations because of the iterative nature of control algorithms that need to predict the behavior of the system and its response to various scenarios of control variables. However, physical simulators of such systems are computationally heavy as they often need to solve non linear equations (e.g. hydraulic equations). Thus, accurate and yet fast numerical models of the network's different components and their interactions are needed. Our proposition to overcome this limitation is the formulation of a numerically efficient and stable surrogate model of DHN simulation. In DHN control optimization, machine learning (ML) was applied at two different levels. The first one focuses on predicting the thermal load [5, 4] and the second, more recent, applying deep reinforcement learning to train autonomous agents to optimally operate DHN [6]. However, to author's knowledge, no attempt to formulate a surrogate model of DHN simulation using ML have been made. In this paper, we present a complete pipeline that was developed for simulating DHN using spatio-temporal graph convolution neural networks (STGCN).

2 PROBLEM STATEMENT

The topology of a DHN can be defined using graph theory where the nodes are either the heating plants or the consumers and the edges are the pipes. In the development of DHN project, there are two major challenges. First a design problem, where the production power is sized with respect to the heat demand of all connected consumers, and the network topology is optimized with regards to heat losses and investment costs. The second, which is the subject of this work, is a control problem. In the latter, given a DHN with a specific topology, the objective is to simulate its dynamic response to different scenarios (a set of exogenous and control variables). More precisely, we want to know the temperatures and mass flow rates at each node of the network.

3 SURROGATE MODELS

During the development phase of this framework, different types of models were evaluated for the studied problem. Here we present the two architectures offering the best performance. The choice of these architectures is the result of a literature study on surrogate models and in particular graph neural networks [1, 7, 3]. The following paragraphs summarize the evaluated architectures.

Encoder-Decoder model: The first architecture that has been evaluated is in the form of an encoder-processor-decoder. The encoder consists of two MLPs that transform the input data and project it into a latent space where each node is assigned two hidden vectors representing respectively its local attributes (heat demand) and also information from global variables updated with neighbors representations. The processor part consists on transforming the nodes hidden representations using three multi-headed attention operations. Finally the decoder is composed of two linear layers that map the nodes representations to the output space.

Recurrent model: The inference problem involves data with temporal dependencies. It is known that recurrent neural networks are suited to this type of data. Therefore, the second architecture that was implemented is a spatio-temporal graph convolution neural network (STGCN) inspired from [7]. In addition to the spatial convolution where the nodes exchange information with their neighbors, this model incorporates a temporal cell based on Gated Recurrent Unit (GRU) in order to update the node representations using their previous hidden states.

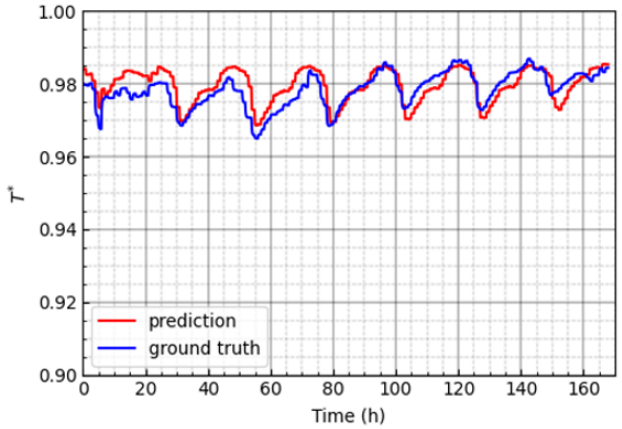


Figure 1: Normalized temperature (T^*) prediction over 1 week for node 13 using Encoder-Decoder model with in the case the supply temperature remains constant.

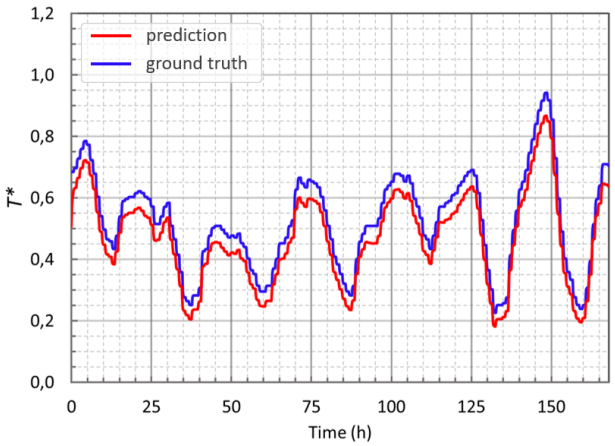


Figure 2: Normalized temperature (T^*) prediction over 1 week for node 13 using STGCN model and with sharp control laws

4 RESULTS AND DISCUSSION

The two models are evaluated against the physical model on an academic study case. The simulation time using the physical model is equal to $t_{sim,4c}$ = 3h2min using a 4-cores CPU processor with a 32-GB RAM. We decided a split of 75% for training and 25% for validation. Thus, 39 weeks are used for training and 13 weeks for validation. Without loss of generality, here we show the predictions for (normalized) return temperatures only :

$$T^* = \frac{T - T_{min}}{T_{max} - T_{min}}$$

First, the models were evaluated on data with low frequency control laws, i.e. supply temperature and the total mass flow rate could remain constant for some time periods. In this case the variable having the strongest weight on the values to be predicted is the heat demand profile. In this case, both models performed very well in predicting the variables of interest. An example is given in Figure 1. It was also found that the RMSE stays below 0.7 Kelvin (0.7K) for all nodes, which is acceptable for the end users, i.e. network operators.

In order to test further the models adaptability and sensitivity to control laws, we tested them with new simulations where the supply temperature curve and the total mass flow rate vary more. Physically, this implied that the return temperature at each node was more affected by the control variables than the heat load itself. The Encoder-Decoder model was not able to capture the correct patterns. On the other hand, the STGCN model captured well enough the system dynamics as shown in Fig. 2. This is the direct result of using GRU to better incorporate the notion of temporality and time dependence into the inference function.

The inference time is different for both models. To simulate one week of operation, the Encoder-Decoder model needs $t_{inf} = 0.019s$ against $t_{sim,4c} = 336s$ for the physical simulation, which accounts for approximately 1.8×10^4 time gain. The STGCN model needs $t_{inf} = 7.71s$, this accounts for approximately 43 time gain.

5 CONCLUSION

In this paper, we presented a work in progress for the application of geometric deep learning as framework for surrogate modeling of district heating networks simulation. The bottleneck to control such systems is their heavy simulations. Therefore, the aim of the surrogate model is to reduce computation time while preserving a high accuracy. Depending on the architecture, the time gain varies in our experiences from a factor of 43 to 1.8×10^4. The next line of work is to deepen the analysis of these models and to fuse more physical constraints during the training phase to respect the physical laws in the predicted values.

ACKNOWLEDGMENTS

This work was supported by the French Ministry of Higher Education and Research.

REFERENCES

[1] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. 2021. Geometric deep learning: grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*.

[2] Annelies Vandermeulen, Bram van der Heijde, and Lieve Helsen. 2018. Controlling district heating and cooling networks to unlock flexibility: a review. *Energy*, 151, 103–115.

[3] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Pietro Romero Adriana and Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

[4] Paul Westermann, Matthias Welzel, and Ralph Evins. 2020. Using a deep temporal convolutional network as a building energy surrogate model that spans multiple climate zones. *Applied Energy*, 278, 115563.

[5] Jiyang Xie, Hailong Li, Zhanyu Ma, Qie Sun, Fredrik Wallin, Zhongwei Si, and Jun Guo. 2017. Analysis of key factors in heat demand prediction with neural networks. *Energy Procedia*, 105, 2965–2970.

[6] Bin Zhang, Amer MYM Ghias, and Zhe Chen. 2022. A double-deck deep reinforcement learning-based energy dispatch strategy for an integrated electricity and district heating system embedded with thermal inertial and operational flexibility. *Energy Reports*, 8, 15067–15080.

[7] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. 2019. T-gcn: a temporal graph convolutional network for traffic prediction. *IEEE transactions on intelligent transportation systems*, 21, 9, 3848–3858.

6 Résumés des articles de démonstration

Computing MAP Inference on Temporal Knowledge Graphs with NeoMaPy

Victor David
Dipartimento di Matematica e Informatica, University of Perugia
Italy
victor.david@unipg.it

Raphaël Fournier-S'niehotta
Conservatoire National des Arts et Métiers
Paris, France
fournier@cnam.fr

Nicolas Travers
Léonard de Vinci Pôle Universitaire, Research Center
Paris, France
nicolas.travers@devinci.fr

1 INTRODUCTION

Markov Logic Networks (MLNs) [7, 12] are a very useful conceptual tool for reasoning over uncertain facts. They combine Markov networks and First Order Logic, by attaching weights to logic formulae. Several MLNs extensions have been devised to work on different types of data [2, 14, 16]. Those uncertain temporal facts generate *conflicts*. Reasoning on those facts often requires to resolve those conflicts. MLNs help find *the most probable state of the world*, gathering a set of facts whose weights have maximal probabilities with a process called *Maximum A-Posteriori* inference (MAP) [10, 11, 13, 15]. However, the state of the art integrating temporal information into MLN is insufficient, and computing the MAP inference usually relies on a heavy data mining process [4].

We have recently introduced an extension of MLNs called Temporal Markov Logic Networks (TMLN) [6], along with a temporal parametric semantics which examined total and partial (in)consistency relations between temporal formulae. Our completely different approach to MAP inference relies on building compatible worlds instead of mining valid worlds.

In this paper, we introduce the NeoMaPy framework, a complete implementation for TMLN reasoning[1]. To achieve this, we extract a conflict graph between facts [1, 9], based on rules and nodes weights. Thus, the MAP inference searches combinations of non-conflict graphs. This approach allows to parameterize MAP inferences with various semantics, computing efficiently with a heuristic and interacting with results for explaining choices of facts.

2 TEMPORAL MARKOV LOGIC NETWORKS

Temporal Markov Logic Networks are based on a Temporal Many-Sorted First-Order Logic TF-FOL which combines formulae and temporal predicates from a temporal domain, to represent temporal facts and rules. The whole formalism of the approach is presented in [5, 6]. TMLN associate a degree of certainty to each formula.

3 THE NEOMAPY APPROACH

The NeoMaPy framework consists of a two-step MAP inference extraction based on a graph database, and a conflict resolution heuristic. This major contribution introduces a parametric, efficient and interactive process.

[1]A companion video is available at https://www.youtube.com/watch?v=c8AzFQMs1I4

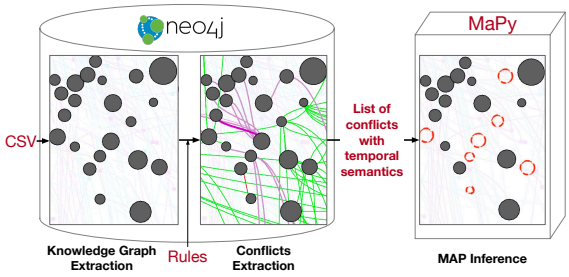


Figure 1: NeoMaPy pipeline for the MAP inference.

Graph of conflicts. This first step transforms a TMLN instantiation into a property graph where constants and predicates become *Concept* nodes with temporal predicates and weights as properties. Rules are expressed as queries on the graph of interactions between *TF* nodes based on their properties, constants and predicates. They produce conflict relationships between *TF* nodes, labelled with a conflict type. *TF* and *Concept* nodes and relationships are stored in a graph database. Thanks to this conflict graph, applying semantics corresponds to a pattern query on the graph, searching for conflicts between *TF* nodes. It reduces the MAP inference to the computation of the maximal subset of consistent *TF* nodes.

Infering the MAP. Once the set of conflictual nodes has been obtained, the MAP inference is computed in two steps: 1) we conduct a pre-processing that structures our data into a set of connected components (*i.e.,* if there is no path between two nodes, they are not connected). 2) for each connected component (*i.e.,* a dictionary) we apply in parallel the MAP inference algorithm MaPy which creates a list of solutions by iteratively trying to add each node to the current solutions. We optimize this process by using a heuristic to eliminate the worst solutions and by restricting the size of this solution list, *i.e.,* by keeping the k best solutions.

4 IMPLEMENTATION

Figure 1 illustrates the architecture of the NeoMaPy framework. The first step extracts the knowledge graph by instantiating facts and ground facts with the Neo4j[2] graph database (nodes' size depends on weights). By applying rules, the graph of conflicts is extracted from facts (green and purple links). The second step exploits the list of conflicts with a parameterized semantics and processes it with the MaPy algorithm implemented in Python (red circles are removed nodes). The source code and datasets are available on GitHub[3].

[2]https://neo4j.com
[3]https://github.com/cedric-cnam/NeoMaPy_Daphne

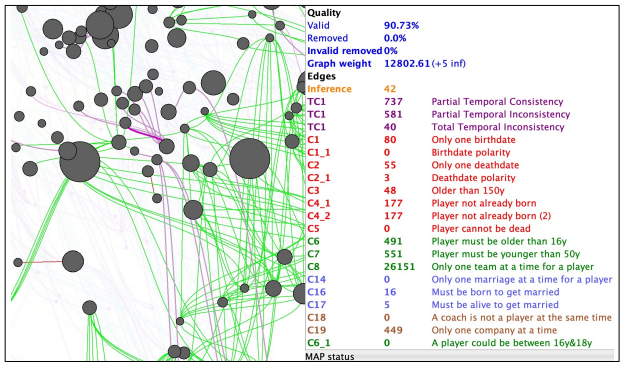


Figure 2: DataViz of the Knowledge graph with conflicts.

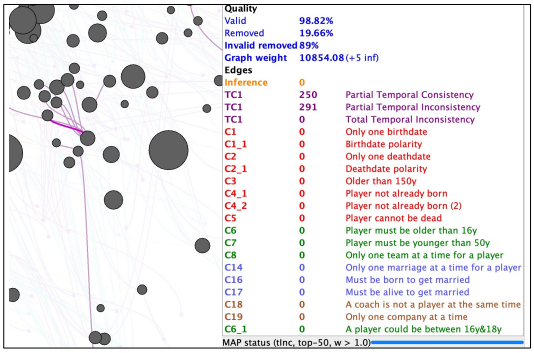


Figure 3: MAP inference DataViz with the Total Temporal Inconsistency semantics.

4.1 MAP Inference Computation

Conflicts extraction from *TF* nodes has been implemented in `Neo4j`. Facts are imported from CSV files. The graph is composed of *Concept* and *TF* nodes. Rules are applied to instantiate ground rules as *Cypher* queries. Conflicts are then instantiated as "*conflict*" relationships on the graph, by searching for `TF` with patterns.

The *Cypher* query below illustrates the generation of conflicts for the *pCon* rule (partial temporal consistency). If two *TF tf1* and *tf2* share the same concepts (*s,o,p*) with opposite polarities (positive or negative information) and a timeframe intersection, it produces a *pCon* conflict between *tf1* and *tf2*. For optimization purposes, concept IDs are repeated in *TF* nodes (*e.g.,* tf1.p = tf2.p). Conflict and inference relationships are typed.

```
MATCH (tf1:TF) -[:s]-> (:Concept) <-[:s]- (tf2:TF)
WHERE tf1.p=tf2.p and tf1.o=tf2.o and tf1.polarity <>
  tf2.polarity AND ( (tf1.date_start <= tf2.date_start
  and tf2.date_start <= tf1.date_end) AND (tf1.date_start
  <= tf2.date_end and tf2.date_end <= tf1.date_end) )
MERGE (tf1)-[c:conflict]-(tf2) SET c.pCon=true;
```

The resulting graph eases the tracability of the MAP inference. Moreover, `MaPy` processes the inference with a parametric semantics extracted with a *Cypher* query.

4.2 Scenarios

A Graphical User Interface was developed using *GraphStream* [4] [8], to improve the reasoning process on uncertain temporal knowledge graphs. The dataset we use contain football facts and rules from [3]. The demonstration will show all `NeoMaPy` steps:

Graph import and conflict extraction. *Concept* and *TF* nodes are imported from CSV files into the *Neo4j* database and inference rules are applied. Then, a set of rules expressed as *Cypher* queries are applied on the graph.

Conflict-graph visualization. The produced conflict graph is visualized as shown in Figure 2. Several interactions are offered to users to explore the knowledge graph, such as node search, clusters of conflict nodes, zooming features. Moreover, graph statistics are computed (graph weights and conflict statistics in Figure 2).

Parametric MAP inferences. Eventually, several MAP inference computations are applied to show the impact on the graph. Simple strategies are compared with different Parametric Temporal Semantics showing the maximization of the argmax $\mathcal{S}(I)$ (*e.g.,* the *tInc total inconsistency* semantics in Figure 3).

REFERENCES

[1] L. Bertossi. *Database Repairs and Consistent Query Answering*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.

[2] Melisachew Wudage Chekol, Jakob Huber, Christian Meilicke, and Heiner Stuckenschmidt. Markov logic networks with numerical constraints. ECAI'16, page 1017–1025, NLD, 2016. IOS Press.

[3] Melisachew Chekol, Giuseppe Pirrò, Joerg Schoenfisch, and Heiner Stuckenschmidt. Marrying uncertainty and time in knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[4] Melisachew Wudage Chekol, Giuseppe Pirro, Joerg Schoenfisch, and Heiner Stuckenschmidt. Tecore: temporal conflict resolution in knowledge graphs. *Proceedings of the VLDB Endowment*, 10:Iss–12, 2017.

[5] Victor David, Raphaël Fournier-S'niehotta, and Nicolas Travers. Neomapy: A framework for computing MAP inference on temporal knowledge graphs. In *IJCAI'23, Macao, China*, pages 7123–7126. ijcai.org, 2023.

[6] Victor David, Raphaël Fournier-S'niehotta, and Nicolas Travers. Neomapy: A parametric framework for reasoning with map inference on temporal markov logic networks. In *CIKM'23*, page 400–409, New York, NY, USA, 2023.

[7] Pedro Domingos and Daniel Lowd. Unifying logical and statistical ai with markov logic. *Communications of the ACM*, 62(7):74–83, 2019.

[8] Antoine Dutot, Frédéric Guinand, Damien Olivier, and Yoann Pigné. GraphStream: A Tool for bridging the gap between Complex Systems and Dynamic Graphs. In *ECCS'07*, Dresden, Germany, October 2007.

[9] Keith W. Hipel, Liping Fang, and D. Marc Kilgour. The graph model for conflict resolution: Reflections on three decades of development. *Group Decision and Negotiation*, 29(1):11–60, 2020.

[10] Feng Niu, Christopher Ré, AnHai Doan, and Jude Shavlik. Tuffy: Scaling up statistical inference in markov logic networks using an rdbms. *arXiv preprint arXiv:1104.3216*, 2011.

[11] Jan Noessner, Mathias Niepert, and Heiner Stuckenschmidt. Rockit: Exploiting parallelism and symmetry for map inference in statistical relational models. In *AAAI'13*, volume 27, pages 739–745, 2013.

[12] Matthew Richardson and Pedro Domingos. Markov Logic Networks. *Machine Learning*, 62(1):107–136, Feb 2006.

[13] Sebastian Riedel. Improving the accuracy and efficiency of map inference for markov logic. *arXiv preprint arXiv:1206.3282*, 2012.

[14] Romain Rincé, Romain Kervarc, and Philippe Leray. Complex event processing under uncertainty using markov chains, constraints, and sampling. In *Rules and Reasoning*, pages 147–163, Cham, 2018. Springer.

[15] Somdeb Sarkhel, Deepak Venugopal, Parag Singla, and Vibhav Gogate. Lifted map inference for markov logic networks. In *Artificial Intelligence and Statistics*, pages 859–867. PMLR, 2014.

[16] Lauro Snidaro, Ingrid Visentini, and Karna Bryan. Fusing uncertain knowledge and evidence for maritime situational awareness via markov logic networks. *Information Fusion*, 21:159–172, 2015.

[4] A Java library for graphs: https://graphstream-project.org/.

More power to SPARQL: From paths to trees

Angelos Christos Anadiotis*
Oracle, Switzerland
angelos.anadiotis@oracle.com

Ioana Manolescu
Inria and IPP, France
ioana.manolescu@inria.fr

Madhulika Mohanty
Inria and IPP, France
madhulika.mohanty@inria.fr

ABSTRACT

Exploring Knowledge Graphs (KGs, in short) to discover facts and links is tedious even for experts with knowledge of SPARQL due to their unfamiliarity with the structure and labels of entities, classes and relations. Some KG applications require finding the connections between groups of nodes, even if users ignore the shape of these connections. However, SPARQL only allows checking if paths exist, not returning them. A recent property graph query language, GPML, allows also returning connecting paths, but not connections between three or more nodes.

We propose to demonstrate RelSearch, a system supporting *extended* SPARQL queries, featuring standard Basic Graph Patterns (BGPs) as well as novel Connecting Tree Patterns (CTPs); each CTP requests *the connections* (paths, or trees) between nodes bound to variables. RelSearch evaluates such extended queries using novel algorithms [1] which, unlike prior keyword search methods, return connections regardless of the edge directions and are independent of how we measure the quality (score) of each connection. We will demonstrate RelSearch's expressivity and efficiency using a variety of RDF graphs, user-selected score functions, and search exploration orders.

REFERENCES

[1] Angelos-Christos Anadiotis, Ioana Manolescu, and Madhulika Mohanty. Integrating connection search in graph queries. In *ICDE*, 2023.

*Work done while at Ecole Polytechnique.

GSTSM Package: Finding Frequent Sequences in Constrained Space and Time

Antonio Castro
Heraldo Borges
antonio.castro@eic.cefet-rj.br
heraldo.borges@eic.cefet-rj.br
CEFET/RJ
Rio de Janeiro, RJ, Brazil

Cassio Souza
Jorge Rodrigues
cassio.souza@eic.cefet-rj.br
jorge.rodrigues@eic.cefet-rj.br
CEFET/RJ
Rio de Janeiro, RJ, Brazil

Fabio Porto
fporto@lncc.br
LNCC
Petrópolis, RJ, Brazil

Esther Pacitti
esther.pacitti@lirmm.fr
INRIA & Univerisity of Montpellier
Montpellier, France

Rafaelli Coutinho
rafaelli.coutinho@cefet-rj.br
CEFET/RJ
Rio de Janeiro, RJ, Brazil

Eduardo Ogasawara
eogasawara@ieee.org
CEFET/RJ
Rio de Janeiro, RJ, Brazil

ABSTRACT

Spatial time-stamped sequences have information about time and space where events occur. Mining such sequences can bring important insights. However, not all sequences are frequent over an entire dataset. Some are only common in subsets of time and space. This article explains the first tool for mining these sequences in constrained space and time: the *GSTSM* R package. It allows users to search for spatio-temporal patterns that are not frequent in the entire database, but are dense in restricted time-space intervals. Thus, making it possible to find non-trivial patterns that would not be found using common data mining tools.

CCS CONCEPTS

• **Information systems → Data mining**.

KEYWORDS

Data Mining, Spatial-Temporal, Time Series, Sequential Mining

1 INTRODUCTION

Data mining tools have been used to find interesting patterns in different areas of knowledge in various problems [1]. The sequence mining knowledge area is a specialization of data mining, focused on finding sequences or series of events in datasets [12, 15]. Such sequences may form patterns, a set of frequent attributes that appear persistently among the dataset. It means that its frequency exceeds a user-defined minimum threshold [3].

Several types of events involve both temporal and spatial data, such as financial to understand sales patterns over time and space [1], and hydrological data for river water quality monitoring in different points over time [1, 2]. They correspond to Time-stamped Sequence (TS) events distributed in space [11]. Mining sequences related to space and time enables to find knowledge related to phenomena that involve both spatial and temporal components, trying to find all sequences of significant, useful, interesting, and non-trivial events [1, 3, 13].

However, spatio-temporal sequential patterns may have low support if considered the entire dataset, but they can be frequent if considered only a period and region [8]. The Generalized Spatial-Time Sequence Miner (*GSTSM*) package can find these patterns, being able to efficiently discover the region and period where they occur. This way, *GSTSM* would be the right tool to find time-localized patterns.

This work describes the process, structure, and usage of the *GSTSM* package using a synthetic (but still complete) example. However, we also provide a glimpse of applicability in a real-world dataset. *GSTSM* was able to found sequential patterns in seismic data. They correspond to seismic horizons, which are important elements in the application domain.

2 RELATED WORK

There are different methods for spatio-temporal data mining. Some use only data mining, searching for frequent patterns, considering only time [10]. Others combine techniques by seeking in time and then grouping in space [7]. Furthermore, there is a diversity in how constraints are handled. Some use global support, a value that is valid for the entire dataset [2]. Others consider local support, using predefined windows of time and space [9].

This work differs by seeking frequent sequences in time that occur in spatial groups. Instead of using predefined constraints for time and space, three density parameters are established: a minimum frequency to be achieved within the period, a maximum distance that a position can be from any other in the group, and a lower limit of distinct positions in the group. Thus, the formalization presented in this work can find different sizes of sequences, time intervals, and spatial regions where a sequence is frequent, based on the concepts of RG, KRG, and SRG introduced in [5].

As far as the conducted research has reached, the only work with a similar approach found in the literature is proposed by [4], which considers one-dimensional space. The present work is a generalization that presents a formalization considering space in its three-dimensional form.

3 DEMONSTRATION OVERVIEW

GSTSM is a package which provides polymorphic functions that let the user extends its functionalities, as it is based on *R* [14] language *S3* classes. The source code can be found at GitHub [6]. The package has a main class named *GSTSM*, that needs the parameters D, P, γ, β, and σ to instantiate an object, explained as follows:

- `D` and `P` represents the TS dataset with their respective positions. Each TS must be associated with one position. It means that the number of timestamped sequences (columns) in `D` must be equal to the number of positions (rows) in `P`.
- for the user defined thresholds values in the range $]0, 1]$ for γ, values starting from 2 for β, and integer values starting from 1 for σ.

A *GSTSM* object is an instance of S3 Class built as a list with all this information. Furthermore, it generates an adjacency matrix that informs each position which other positions are at a maximum distance of σ.

The *GSTSM* package has the *mine()* method that implements the entire process of finding frequent sequences. It receives as input a *GSTSM* object and provides as output a list of the SRGs of all sizes found. The user does not need to call any other method to get the results. This method calls and passes all the necessary parameters to make the entire process transparent to the user.

The other methods used in each process step are polymorphic and can be extended by the user. It gives the user the ability to try its implementation. These are described as follows:

- *find()* has two input parameters: a *GSTSM* object and a set of candidate sequences of size k. It provides as output the KRGs for each candidate.
- *merge()* has also two input parameters: a *GSTSM* object and a set of candidate sequences of size k containing information about the KRG of each one. The method returns the SRGs with the candidate sequences of size k.
- *generate_candidates()* has two input parameters: a *GSTSM* object and a set of SRGs of size k. There are no SRGs to pass to generate candidates of size one, a `NULL` value can be used. The method provides the candidate sequences of size $k + 1$.

An illustrative example shows the use of the *GSTSM* package functions. To start, the first action is installing and loading *GSTSM* package and then setting all the inputs for the package: D, P, γ, β, and σ. For D, we use a simple dataset. For P, positions in a row are used, with one unit distance. Each position is associated with a time series, such as p_1 to t_1 and p_2 to t_2. The values for the user-defined thresholds are: $\gamma = 0.8$, $\beta = 2$, and $\sigma = 1$. After setting the input parameters, we can instantiate the *GSTSM* object and execute the *mine()* method. Listing 1 shows the code using the R command line.

Listing 1: R example of the use of *GSTSM*

```
# loading the GSTSM package
 > library("gstsm")
# loading Spatial Timestamped Sequence
 > path <-
   "https://eic.cefet-rj.br/~dal/wp-content/uploads/2023/05/"
 > load(url(paste(path, "dataset.rdata"))) # dataset D
 > load(url(paste(path, "positions.rdata"))) # positions P
# mining dataset
 > gstsm_object <- gstsm(D, P, gamma=0.8, beta=2, sigma=1)
 > result <- mine(gstsm_object)
```

4 CONCLUSION

GSTSM is the first tool for mining sequences in spatial time-stamped sequences datasets able to discover constrained patterns in time and space with all three dimensions. The package discovers patterns that may not be frequent over an entire dataset but are grouped in space and frequent in a time interval. It would not be easy to find these patterns without this tool. The results can differ from conventional data mining tools and give different insights about data behavior. The patterns are groups of positions and periods where the sequences are frequent according to the input parameters. The package is also extensible, enabling users to incorporate heuristics and optimizations to drive the discovery of patterns.

ACKNOWLEDGMENTS

The authors thank CNPq, CAPES, and FAPERJ for partially sponsoring this research.

REFERENCES

[1] H. Alatrista-Salas, J. Azé, S. Bringay, F. Cernesson, N. Selmaoui-Folcher, and M. Teisseire. 2015. A knowledge discovery process for spatiotemporal data: Application to river water quality monitoring. *Ecological Informatics* 26, P2 (2015), 127–139. https://doi.org/10.1016/j.ecoinf.2014.05.011

[2] H. Alatrista-Salas, S. Bringay, F. Flouvat, N. Selmaoui-Folcher, and M. Teisseire. 2016. Spatio-sequential patterns mining: Beyond the boundaries. *Intelligent Data Analysis* 20, 2 (2016), 293–316. https://doi.org/10.3233/IDA-160806

[3] B. Aydin and R.A. Angryk. 2016. Spatiotemporal event sequence mining from evolving regions. In *Proceedings - International Conference on Pattern Recognition*, Vol. 0. 4172–4177. https://doi.org/10.1109/ICPR.2016.7900288

[4] R. Campisano, H. Borges, F. Porto, F. Perosi, E. Pacitti, F. Masseglia, and E. Ogasawara. 2018. Discovering tight space-time sequences. In *Lecture Notes in Computer Science*, Vol. 11031. 247–257. https://doi.org/10.1007/978-3-319-98539-8_19

[5] A. Castro, H. Borges, R. Campisano, E. Pacitti, F.Porto, R. Coutinho, and E. Ogasawara. 2021. Generalização de Mineração de Sequências Restritas no Espaço e no Tempo. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*. SBC, 313–318. https://doi.org/10.5753/sbbd.2021.17891

[6] A. Castro, C. Souza, J. Rodrigues, E. Pacitti, F. Masseglia, R. Coutinho, and E. Ogasawara. 2022. *GSTSM Package*. Technical Report. https://cran.rstudio.com/web/packages/gstsm/index.html, CRAN.

[7] C. Flamand, M. Fabregue, S. Bringay, V. Ardillon, P. Quénel, J.C. Desenclos, and M. Teisseire. 2014. Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana. *Journal of the American Medical Informatics Association : JAMIA* 21, e2 (2014), e232–240. https://doi.org/10.1136/amiajnl-2013-002348

[8] Y. Huang, L. Zhang, and P. Zhang. 2008. A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and Data Engineering* 20, 4 (2008), 433–448. https://doi.org/10.1109/TKDE.2007.190712

[9] B. Koseoglu, E. Kaya, S. Balcisoy, and B. Bozkaya. 2020. ST Sequence Miner: visualization and mining of spatio-temporal event sequences. *Visual Computer* 36, 10-12 (2020), 2369–2381. https://doi.org/10.1007/s00371-020-01894-6

[10] K. Li and Y. Fu. 2014. Prediction of human activity by discovering temporal sequence patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 8 (2014), 1644–1657. https://doi.org/10.1109/TPAMI.2013.2297321

[11] C.H. Mooney and J.F. Roddick. 2013. Sequential pattern mining - Approaches and algorithms. *Comput. Surveys* 45, 2 (2013). https://doi.org/10.1145/2431211.2431218

[12] S. Parthasarathy, M.J. Zaki, M. Ogihara, and S. Dwarkadas. 1999. Incremental and interactive sequence mining. In *International Conference on Information and Knowledge Management, Proceedings*. 251–258. https://doi.org/10.1145/319950.320010

[13] G. Sunitha and A. Rama Mohan Reddy. 2014. Mining frequent patterns from spatiotemporal data sets: A survey. *Journal of Theoretical and Applied Information Technology* 68, 2 (2014), 265–274.

[14] R. Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Technical Report. https://www.R-project.org/, Vienna, Austria.

[15] Mohammed J. Zaki. 2000. Sequence mining in categorical domains: incorporating constraints. In *Proceedings of the ninth international conference on Information and knowledge management (CIKM '00)*. Association for Computing Machinery, New York, NY, USA, 422–429. https://doi.org/10.1145/354756.354849

InteGraal: a Tool for Data-Integration and Reasoning on Heterogeneous and Federated Sources

Jean-François Baget, Pierre Bisquert, Michel Leclère, Marie-Laure Mugnier,
Guillaume Pérution-Kihli, Florent Tornil, Federico Ulliana
first.last@inria.fr
LIRMM, Inria, Univ Montpellier, CNRS, INRAE
Montpellier, France

ABSTRACT

We propose to demonstrate InteGraal, a tool for reasoning over heterogeneous and federated data sources. InteGraal is a highly modular tool constituted by two main components. The first is the data-integration layer which allows the users to build a federated factbase over a collection of sources. The second is the automated reasoning layer, which provides powerful means for the declarative exploitation of data through the expressive formalism of *existential rules*. This demonstration proposes to showcase the use of the tool in use-cases of data exploitation as well as to present its architecture and its dedicated query answering mechanisms.

ACKNOWLEDGMENTS

We are grateful to Julien Cufi and Patrice Buche for their help setting up the demonstration scenario. This work was partially supported by the Inria-DFKI project R4Agri and the ANR project CQFD (ANR-18-CE23-0003).

PathWays: entity-focused exploration of heterogeneous data graphs

Nelly Barret
nelly.barret@inria.fr
Inria and Institut Polytechnique de Paris, France

Antoine Gauquier
antoine.gauquier@etu.imt-nord-europe.fr
Institut Mines Télécom, France

Jia Jean Law
jia-jean.law@polytechnique.edu
Ecole Polytechnique, France

Ioana Manolescu
ioana.manolescu@inria.fr
Inria and Institut Polytechnique de Paris, France

ABSTRACT

Graphs, and notably RDF graphs, are a prominent way of sharing data. As data usage democratizes, users need help figuring out the useful content of a graph dataset. In particular, journalists with whom we collaborate [1] are interested in identifying, in a graph, the *connections between entities*, e.g., people, organizations, emails, etc.

We propose PathWays, an interactive tool for exploring data graphs through *their data paths connecting Named Entities (NEs, in short)*; each data path leads to a tabular-looking set of results. NEs are extracted from the data through dedicated Information Extraction modules. PathWays leverages the ConnectionLens platform [2, 4] and follow-up work on dataset abstraction [3]. Its novelty lies in its interactive and efficient approach to enumerate, compute, and analyze NE paths.

REFERENCES

[1] Anadiotis, A., Balalau, O., Bouganim, T., et al. Empowering investigative journalism with graph-based heterogeneous data management. *IEEE DEBull.* (2021).
[2] Anadiotis, A., Balalau, O., Conceicao, C., et al. Graph integration of structured, semistructured and unstructured data for data journalism. *Inf. Systems 104* (2022).
[3] Barret, N., Manolescu, I., and Upadhyay, P. Abstra: toward generic abstractions for data of any model (demonstration). In *CIKM* (2022).
[4] Chanial, C., Dziri, R., Galhardas, H., et al. ConnectionLens: Finding connections across heterogeneous data sources (demonstration). *PVLDB 11*, 12 (2018).

Interpretable Clustering of Multivariate Time Series with Time2Feat

Angela Bonifati
Lyon 1 University, Liris CNRS
Lyon, France
angela.bonifati@univ-lyon1.fr

Francesco Del Buono
University of Modena and Reggio Emilia
Modena, Italy
francesco.delbuono@unimore.it

Francesco Guerra
University of Modena and Reggio Emilia
Modena, Italy
francesco.guerra@unimore.it

Miki Lombardi
Adobe
Paris, France
mlombardi@adobe.com

Donato Tiano
University of Modena and Reggio Emilia
Modena, Italy
donato.tiano@unimore.it

CCS Concepts: • **Computing methodologies** → **Cluster analysis**; **Semi-supervised learning settings**.

Keywords: Clustering, Features Selection, Semi Supervised Clustering, Interpretability

1 Introduction

A Multivariate Time Series (MTS) is a collection of multiple univariate time series (signals) that are observed simultaneously over time and provide insight into time-dependent phenomena. MTS analysis enables the examination of variable relationships over time, offering important insights into underlying phenomena. This can be useful for modeling complex systems, making data-driven decisions, and improving efficiency and productivity across various domains, such as finance and economics, environmental science, healthcare, and social science. MTS analytics involves both supervised and unsupervised techniques, which cover a range of tasks, including classification, clustering, pattern discovery, and forecasting. Among these, clustering analysis has become popular in applications where sensors generate large amounts of data.

While there has been a lot of research on clustering techniques for univariate time series (UTS), the field of clustering multivariate time series is still in its early stages. Proposals adapt clustering approaches designed for UTS to MTS after applying dimensionality reduction techniques. Examples of such techniques are based on the Principal Component Analysis (PCA), which enables the conversion of a set of correlated features in the high dimensional space into a set of uncorrelated features in the low dimensional space. Nevertheless, the resulting clusters suffer from poor explainability as the original dimensions are lost. More recently, approaches based on Deep Neural Networks (DNNs), and in particular Variational Autoencoders (VAEs) have been used to generate MTS encodings before applying clustering methods. Although these solutions might exhibit high performance, the resulting clusters are based on latent dimensions that remain unexplainable to the end-users.

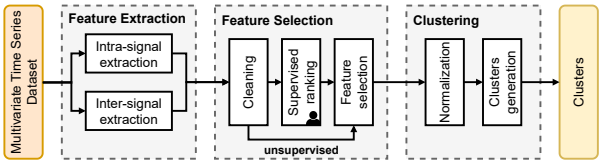


Figure 1. The Time2Feat pipeline.

In this paper, we demonstrate Time2Feat[1, 2], the first system that deals with explainable results of Multivariate Time Series clustering using an end-to-end feature-based pipeline. The pipeline applies clustering techniques to interpretable features automatically extracted from the signals composing the MTS. While the pipeline can be executed without any user interaction (i.e., running the *unsupervised mode*), Time2Feat can also incorporate user annotations on small dataset samples to select generated features. This kind of user's *semi-supervision* can improve the accuracy of results and the quality of explanations.

Through the demonstration, users can experiment with the end-to-end process implemented in the pipeline to generate clusters and explanations from multiple MTS datasets using both unsupervised and semi-supervised modes.

2 The Time2Feat system

The Time2Feat data analysis pipeline is comprised of three distinct components: feature extraction, feature selection,

and clustering, as depicted in Figure 1. The input to the pipeline is a Multivariate Time Series, along with the number of desired clusters. If the number of clusters is not explicitly specified, the pipeline can determine it using a heuristic approach.

There are two operational modes for Time2Feat, unsupervised mode and semi-supervised mode. In unsupervised mode, the pipeline requires no additional input beyond the Multivariate Time Series and the number of desired clusters. The pipeline automatically extracts relevant features, selects the most important ones, and performs clustering to identify the number of specified clusters.

In semi-supervised mode, the user has the option of providing a subset of labeled samples to improve the accuracy of the clustering process. This information is utilized to fine-tune the feature selection and clustering steps, ultimately resulting in improved clustering accuracy.

Feature Extraction. The goal of this component is to generate a comprehensive and detailed representation of the MTS in the dataset via the extraction of a large spectrum of features, each describing the signals composing the MTS in isolation or pairs. *Intra-signal Features* are extracted by applying the library `tsfresh`, which can extract 700+ features encoding the signal description from the perspective offered by a specific analysis method, such as Distribution Analysis, Statistical Analysis, etc. `tsfresh` extracts features that are interpretable for users who know how the statistical measure summarizes the time series values. Moreover, Time2Feat captures *inter-signal relationships* by quantifying the relatedness between pairs of signals using eight metrics, such as correlation and Euclidean distance. All of the features extracted via this process represent the primary characteristics of the time series. These features are entirely based on statistical calculations, which results in them being fully explainable and interpretable.

Features Extraction at work. The RacketSports dataset[1] describes four kinds of shots performed by people playing badminton or squash. Two sensors gather data in a three-dimensional space, thus producing three signals per sensor. Suppose we are asked to analyze the dataset and no details on the activities that MTS describes are provided to us. Time2Feat through the Feature Extraction component, generates 4722 intra-signal and 120 inter-signal features to describe the dataset. While individually interpretable, the sheer number of features can produce noise in the generation of the clusters and cannot be managed by users for their interpretation.

Feature Selection. The feature extraction process generates a vast number of features, making it necessary to reduce its dimensionality to improve interpretability and clustering performance. If Time2Feat is running with the *semi-supervised mode*, the available labels are used to rank the relevance of the features for identifying a subset capable of generating clusters via an ANOVA based analysis. Irrespective of the availability of labels, the Principal Feature Analysis (PFA) technique is applied to select the most meaningful features. The PFA technique is chosen because it not only ensures conciseness but also promotes diversity by selecting the principal features that retain the maximum variability of the features in the lower-dimensional space.

Features Selection at work. The Feature Selection component reduces the number of features to 142 intra-signal and 11 inter-signal features when running in the semi-supervised mode. If the user can provide labels for at least 20% of the MTS, the number of features is reduced to 12 intra-signal and 3 inter-signal features.

Clustering. Time2Feat includes three techniques (Hierarchical, KMeans, Spectral) for generating the clusters. The hierarchical clustering technique is selected as default, since it achieved the best accuracy in our experiments. The Time2Feat system leverages state-of-the-art heuristics (e.g., applying the well-known Elbow method) or user preferences to select the number of clusters to generate. The clusters are wholly derived from the previously selected features. Given the interpretability of these features, it is feasible to analyze the resulting clusters and gain an understanding of the factors that influenced their creation.

Clustering Component at work. If the user selects to generate 4 clusters and the system runs in the unsupervised mode, the quality of the result measured with the Adjusted Mutual Information (AMI) is around 0.35. The reason for the low quality result (that in any case overcomes the competing approaches) is mainly due to the complexity of the problem. By running the semi-supervised mode and providing 20% of labels, the quality of the clusters improves to 0.56. If the user selects to generate 2 clusters, Time2Feat is able to better discriminate between the activities. The AMI obtained with the unsupervised approach is 0.77 and the one with the semi-supervised approach is 0.86.

References

[1] Angela Bonifati, Francesco Del Buono, Francesco Guerra, Miki Lombardi, and Donato Tiano. 2023. Interpretable Clustering of Multivariate Time Series with Time2Feat. *Proceedings of the VLDB Endowment* 16, 12 (2023), 3994–3997.

[2] Angela Bonifati, Francesco Del Buono, Francesco Guerra, and Donato Tiano. 2022. Time2Feat: learning interpretable representations for multivariate time series clustering. *Proceedings of the VLDB Endowment* 16, 2 (2022), 193–201.

[1] http://www.timeseriesclassification.com/description.php?Dataset=RacketSports

HEADWORK: Powering the Crowd with Tuple Artifacts

David Gross-Amblard
first.last@irisa.fr
Univ Rennes / Irisa Lab
France

Marion Tommasi
Inria Lille-Nord Europe
Université de Lille
France

Iandry Rakotoniaina
UMR 7204, MNHN-CNRS-UPMC,
CESCO, Paris
France

Constance Thierry
first.last@irisa.fr
Univ Rennes / Irisa Lab
France

Rituraj Singh
Univ Rennes / Irisa Lab
France

Leo Jacoboni
Wirk
France

ABSTRACT

In this demo we introduce HEADWORK, an open-source academic platform for the crowdsourcing of complex tasks. Besides classical crowdsourcing features, HEADWORK eases the development of crowdsourcing campaigns through a full relational abstraction of relevant concepts (participants, skills, tasks, current answers, decision procedures, GUI, etc.). It allows the orchestration of complex dynamic tasks using so-called *tuple artifacts* (i.e. finite-state automata which transition guards and actions are SQL-defined, on an evolving database). The demo will illustrates these key features, both from the participant and developer point of view.

1 INTRODUCTION

Crowdsourcing is a technique to solve tasks by automatically asking questions to humans. Successful examples are Zooniverse [8], Foldit[3] for participative science, and Amazon Mechanical Turk for rewarded tasks. At the core of crowdsourcing platforms are *micro-tasks*: simple questions awaiting for a simple answer. But while the crowdsourcing of micro-tasks is well studied, recent works turn their attention to *macro-tasks* [5], that require a chain of interactions with humans, using various steps and intermediate decisions. Several systems has been considered to handle this kind of tasks [1, 6], but they rely on a low-level, procedural description of interactions.

In this demo, we present HEADWORK, a ready-to-use, academic crowdsourcing platform for the deployment of complex tasks. In order to limit the task designer's efforts, the HEADWORK platform proposes a full relational abstraction of relevant crowd concepts and algorithms. The orchestration of macro-tasks is realized through *tuple artifacts* [4], that are finite state automata operating on a database, which transitions are guarded by SQL conditions and which trigger SQL actions.

To promote the adoption of HEADWORK, the platform is fully open-source (AGPL), and a demo server is available[1]. The platform has already been used for participative science campaigns, and is compatible with rewarded crowdsourcing.

[1] https://headwork.irisa.fr

2 MODEL

The HEADWORK platform is data-oriented. Our goal is to focus on transforming data from the crowd rather than dealing with low-level programming issues. We illustrate below our relational abstraction, the template mechanism, and explain the deployment of micro and macro-tasks.

2.1 Relational Abstraction

Several built-in tables are available. Basically:

- `user` : gathers information about crowd participants;
- `skill` : contains skill definitions (as keywords and levels of expertise), used for tasks and user profiles;
- `template` : provides classical user interactions (expressed in HTML and Javascript);
- `task` : contains the questions for the crowd;
- `profile` : allows to specify which skill is relevant for a task;
- `answer` : saves participant contributions and intermediate computations.

Micro-tasks are then built on these notions.

2.2 Micro-Tasks

HEADWORK comes with different language flavour. A domain-specific language that we call *Crowdy* is available, allowing to express simply a wide variety of micro-tasks. If needed, task designers have full control of the SQL counterpart. SQL expressions can also be used in specific Crowdy statements. It is noteworthy that letting the task designer access to a full SQL engine is a potential security threat. We will come back to this question in the next section.

2.3 Template Mechanism

HEADWORK features an extensible template mechanism, that allows the task designer to re-use typical crowd interactions, but also to propose new ones to the community. Basic templates are classical HTML form-like inputs such as text, text area, lists and radio buttons. More sophisticated templates such as selectors for geographical maps, image selectors, audio/video playing and audio recording are also available.

The general architecture of a template is an HTML snippet whose interaction is driven by a Javascript code. The code can contain text tags that are populated by a Crowdy statement. The only constraint is to provide the output as a specific field in JSON format, so that HEADWORK is able to process it into the `answer` table.

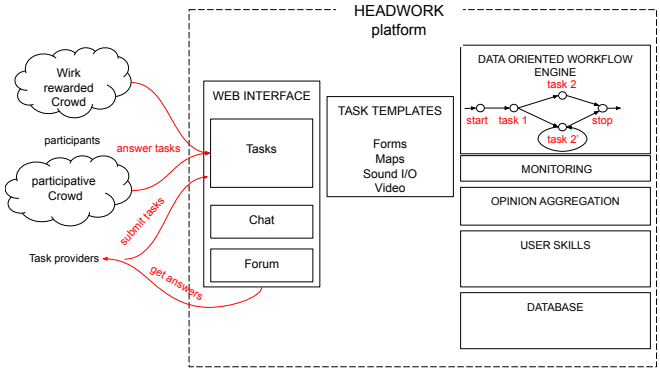


Figure 1: The HEADWORK architecture

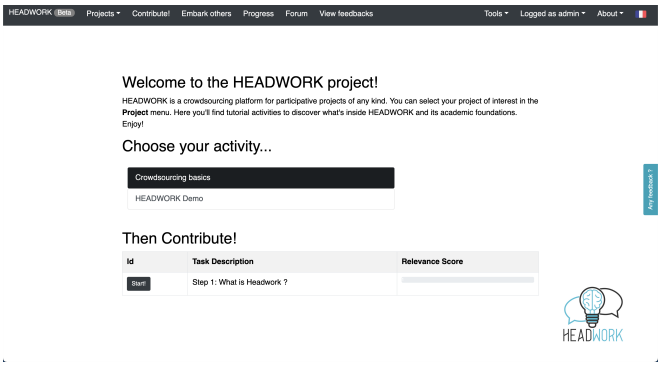


Figure 2: HEADWORK main interface, with projects, activities and tasks, ordered by relevance wrt. participant's skills

2.4 Macro-Tasks

In short, macro-tasks are workflows of simple tasks, which order and content can evolve according to participant answers and crowd decisions. In HEADWORK, a macro-task is driven by a *tuple artifact* : a finite state automaton which transition conditions (guards) and actions are expressed in Crowdy (hence SQL at a low level). Generally speaking, a transition in a tuple artifact has the following structure:

$$\text{state } s \xrightarrow[\text{actions: } \alpha]{\text{guard: } \gamma} \text{state } s',$$

meaning that, if we are in state s with database DB, and the guard query $\gamma(DB)$ is true, then we go to state s', with the new database $\alpha(DB)$.

Since guards and actions can be defined completely with queries on the HEADWORK relational schema, and since any number of states can be envisioned, a wide set of task compositions can be expressed: sequences of questions, conditional branching, loops. Computations and aggregations benefit from the full power of SQL, extended with crowd-style operators such as majority voting. Specific cohort of participants can be defined thanks to queries on the skill and profile tables.

3 THE HEADWORK PLATFORM

The platform is organized as follows (Figure 1). Task providers submit a job as a JSON file encoding the crowd data oriented workflow, in the SQL or Crowdy language, based on the various available templates. The workflow engine (written in PHP hosted by Apache) then processes the automaton and render tasks to participants through the Web interface (Javascript, Bootstrap, Figure 2). Participants can create an account, give their profile (skills), see the list of available tasks ranked according to their skills, and start contributing. All information are available in a Mysql database. If required, HEADWORK is compatible with a rewarded pool of participants through the Wirk service, to speed-up macro-tasks that could not wait for benevolent participants.

4 FUTUR WORK

Our hope is to make HEADWORK an academic laboratory for studies in macro-task crowdsourcing, while hosting real participative and citizen science projects. In the short future we plan to implement richer, hierarchical skill models [7] and to allow for automatic workflow verification [2].

ACKNOWLEDGEMENTS

We would like to thank the numerous interns from Rennes University that contributed to pieces of this project. This work was also partially funded by the French National Research Agency (ANR) grant HEADWORK[2] (ANR-16-CE23-0015).

REFERENCES

[1] Salman Ahmad, Alexis Battle, Zahan Malkani, and Sepander Kamvar. The jabberwocky programming environment for structured social computing. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 53–64, New York, NY, USA, 2011. ACM.

[2] Pierre Bourhis, Loïc Hélouët, Zoltán Miklós, and Rituraj Singh. Data centric workflows for crowdsourcing. In Ryszard Janicki, Natalia Sidorova, and Thomas Chatain, editors, *Application and Theory of Petri Nets and Concurrency - 41st International Conference, PETRI NETS 2020, Paris, France, June 24-25, 2020, Proceedings*, volume 12152 of *Lecture Notes in Computer Science*, pages 24–45. Springer, 2020.

[3] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit players. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 2010.

[4] Alin Deutsch, Richard Hull, Fabio Patrizi, and Victor Vianu. Automatic verification of data-centric business processes. In *Proceedings of the 12th International Conference on Database Theory*, ICDT '09, page 252–267, New York, NY, USA, 2009. Association for Computing Machinery.

[5] Vassillis-Javed Khan, Konstantinos Papangelis, Ioanna Lykourentzou, and Panos Markopoulos. Macrotask crowdsourcing. *Cham: Springer International Publishing*, 2019.

[6] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. Turkit: Human computation algorithms on mechanical turk. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 57–66, New York, NY, USA, 2010. ACM.

[7] Panagiotis Mavridis, David Gross-Amblard, and Zoltán Miklós. Using hierarchical skills for optimized task assignment in knowledge-intensive crowdsourcing. In *Proceedings of the 25th International Conference on World Wide Web*, pages 843–853. International World Wide Web Conferences Steering Committee, 2016.

[8] Robert Simpson, Kevin R Page, and David De Roure. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd international conference on world wide web*, pages 1049–1054, 2014.

[2] https://headwork.irisa.fr/headwork-web/

qEndpoint: A Wikidata SPARQL endpoint on commodity hardware

Antoine Willerval
antoine.willerval@etu.univ-lyon1.fr
Lyon 1 University, CNRS Liris
Lyon, France
The QA Company SAS
Saint Etienne, France

Dennis Diefenbach
dennis.diefenbach@the-qa-company.com
The QA Company SAS
Saint Etienne, France

Angela Bonifati
angela.bonifati@univ-lyon1.fr
CNRS, LIRIS UMR 5205, Lyon 1 University
Lyon, France

ABSTRACT

In this work, we demonstrate how to setup a Wikidata SPARQL endpoint on commodity hardware resources. We achieve this by using a novel triple store called qEndpoint, which uses a read-only partition based on HDT and a write partition based on RDF4J. We show that qEndpoint can index and query the entire Wikidata dump (currently 17 billion triples) on a machine with 600GB SSD, 10 cores and 10GB of RAM, while keeping the query performance comparable with other SPARQL endpoints indexing Wikidata.
In a nutshell, we present the first SPARQL endpoint over Wikidata that can run on commodity hardware while preserving the query run time of existing implementations. Our work goes in the direction of democratizing the access to Wikidata as well as to other large-scale Knowledge Graphs published on the Web. The source code of qEndpoint along with the query workloads are publicly available.

CCS CONCEPTS

• **Information systems → DBMS engine architectures**.

KEYWORDS

qEndpoint, HDT, Wikidata, SPARQL

System	Loading Time	#Triples	RAM	Index size	Doc
Apache Jena	9d 21h	13.8 B	64 GB	2TB	1
Virtuoso	several days[1] (preprocessing) + 10h	11.9 B	378 GB	NA	2
Blazegraph	~5.5d	11.9 B	128 GB	1.1 T	3
Stardog	9.5 h	16.7 B	256 GB	NA	4
QLever	14.3 h	17 B	128 GB	823 GB	5
qEndpoint	50 h	17.4 B	10 GB	294 GB	6

Table 1: Wikidata Index characteristics for different endpoints

(1) https://wiki.bitplan.com/index.php/WikiData_Import_2020-08-15
(2) https://community.openlinksw.com/t/loading-wikidata-into-virtuoso-open-source-or-enterprise-edition/2717
(3) https://addshore.com/2019/10/your-own-wikidata-query-service-with-no-limits/
(4) https://www.stardog.com/labs/blog/wikidata-in-stardog/
(5) https://github.com/ad-freiburg/qlever/wiki/Using-QLever-for-Wikidata
(6) https://github.com/the-qa-company/qEndpoint/wiki/Use-qEndpoint-to-index-a-dataset

Task	Time	Description
Dataset download	7 h	Download the dataset[2]
HDT compression	45 h	Creating HDT
HDT co-index gen	5 h	Creating OPS/PSO/POS indexes
Loading the index	2 min	Start the endpoint

Table 2: Time split during the loading of the Wikidata dataset.

File name	File size	Usage
index_dev.hdt	183GB	Dictionary + SPO index
index_dev.hdt.index.v1-1	113GB	OPS/PSO/POS indexes
native-store	16KB	RDF4J store
qendpoint.jar	82MB	qEndpoint

Table 3: Sizes of each components of qEndpoint (total: 296GB)

Received 3 February 2023; revised ??; accepted ??

7 Résumés des articles de doctorant

Representation Learning for Relational Structures

Alan Gany*
alan.gany@universite-paris-saclay.fr
CNRS-LISN
Paris, France

ABSTRACT

Over the last decades, important improvements in deep learning based methods have been made, allowing to learn rich and versatile representations for text, graphs, or images. In the data management area, an important question is to explore to what extent these methods may be useful to represent relational structures, which may be incomplete (e.g., with missing values) or uncertain. Indeed, many recent research works focus on reconstructing or extracting data from relational structures using learned embeddings, for either tuple classification, entity resolution, or other data cleaning tasks. In this work, we intend to pursue in this research direction, focusing on the important problem of *missing value imputation*.

Our aim is extend the state of the art for this problem, which focused mainly on machine learning datasets consisting of a single relation, to general and practically relevant relational structure. Concretely, we will build a relational structure representation that can capture both syntactic and semantic similarities and correlations, capturing all data facets, in order to have the most effective embeddings.

Our proposal is to construct a tripartite graph representation of the relational data, including the attribute values, the attribute names, and the tuples of the relational structure. This graph captures constraints such primary / foreign key, as well as inclusion dependencies.

The subsequent objective is then to generate embeddings for the nodes of the graph and, based on them, to retrieve for each missing value the most likely candidate one from the same domain. For this embedding task, we have used either task agnostic models such as random walks or tasks-specific deep graph learning models. As we are mapping our relation problem to a graph one, we must ensure that the generated embeddings retain properties such as isomorphism between the different graphs that can be built from the same relational structure. Relations are order (permutation) invariant, and should also transfer to the corresponding graph representations thereof.

Finally, we propose two data imputation methods that can be applied to relation schemas consisting of several interconnected relations. The first method exploits the geometric properties of the generated embeddings to fill each missing value of the database. The second method exploits the task specific deep learning model in order to find candidates for the missing values. Both of these models have the advantage to work on tuples having more than one missing element, without being specific to a single attribute. Moreover, because the graph we build can capture foreign / primary key constraints, the methods proposed allow to retrieve candidate values across different relations.

KEYWORDS

Representation Learning, Graph Learning, Relational Structures

Graph versioning for evolving urban data
Versionnement de graphe pour les données urbaines évolutives

Jey Puget Gil
jey.puget-gil@liris.cnrs.fr
Université de Lyon, Université Claude Bernard, LIRIS, UMR-CNRS 5205
Villeurbanne, FRANCE

Emmanuel Coquery
emmanuel.coquery@univ-lyon1.fr
Université Claude Bernard, LIRIS, UMR-CNRS 5205
Villeurbanne, FRANCE

John Samuel
john.samuel@cpe.fr
Université de Lyon, CPE Lyon, LIRIS, UMR-CNRS 5205
Villeurbanne, FRANCE

Gilles Gesquière
gilles.gesquiere@univ-lyon2.fr
Université de Lyon, Université Lumière Lyon 2, LIRIS, UMR-CNRS 5205
Villeurbanne, FRANCE

ABSTRACT

The continuous evolution of cities poses significant challenges in terms of managing and understanding their complex dynamics. With the increasing demand for transparency and the growing availability of open urban data, it has become important to ensure the reproducibility of scientific research and computations in urban planning. To understand past decisions and other possible scenarios, we require solutions that go beyond the management of urban knowledge graphs. In this work, we explore existing solutions and their limits and explain the need and possible approaches for querying across multiple graph versions.

RÉSUMÉ

L'évolution continue des villes pose des défis importants en termes de gestion et de compréhension de leurs dynamiques complexes. Avec la demande croissante de transparence et la disponibilité grandissante de données urbaines ouvertes, il est devenu important d'assurer la reproductibilité de la recherche scientifique et des calculs dans le domaine de l'urbanisme. Pour comprendre les décisions passées et d'autres scénarios possibles, nous avons besoin de solutions qui vont au-delà de la gestion des graphes de connaissances urbaines. Dans ce travail, nous explorons les solutions existantes et leurs limites, et expliquons le besoin et les approches possibles pour l'interrogation à travers de multiples versions de graphes.

CCS CONCEPTS

• **Information systems** → *Geographic information systems*; **Resource Description Framework (RDF)**; **Web Ontology Language (OWL)**.

KEYWORDS

RDF, versioning, graph, urban data, deduction

MOTS CLÉS

RDF, versionnement, graphe, données urbaines, déduction

1 INTRODUCTION AND MOTIVATION

Urban planners, historians, archaeologists, and researchers are continuously analyzing the constant development of cities. They are interested in an understanding of the possible versions and scenarios of the city [2], both in the past and in the future, if certain decisions were to be made. The choices made and the lessons learned in urban planning in the past serve as a guide for future decisions.

As the availability of open data increases across all sectors, so too does the demand for transparency in decision-making. Urban data come in different forms, they can be structured (sensors, building data, ...), semi-structured (urban system logs, ...) or unstructured (images, text, ...). Decisions are made on the basis of the data available at a given point in time. In other words, both the most recent version of the city and the previous versions are taken into account. In certain cases, complex calculations on this existing data guide the decision-makers. Reproducibility of these calculations is therefore also a requirement for transparency.

We provide the following sample queries from urban planning project proposals to better illustrate our research work:

- **Q1**: *Which city versions have a metro station accessible to people with disabilities?*
- **Q2**: *Across multiple concurrent city versions, what is the maximum known height of a particular building?* (aggregation)

Our previous research work proposed the use of graph formats [4] for the transformation and management of heterogeneous and concurrent urban data. In this work, we want to go beyond this and explore and develop a system that can query multiple versions of the graph simultaneously to answer complex queries like the ones above. This requires not only versioning of code (complex calculations) and data. It also requires efficient querying techniques across versions.

This article briefly reviews different ways to address the need for effective tools and methodologies to analyze urban development, emphasizing the importance of versioned data management.

2 STATE OF THE ART

Data and code evolution have been at the heart of many recent research and industrial advances. Taken together, they make up an important part of urban knowledge evolution. Given their growing use, our research is particularly focused on version control systems.

2.1 Code and Data versioning

Versioned repositories are systems that track and manage changes to data and code over time, allowing researchers to maintain a historical record of their work and facilitate collaboration. When these two concepts are integrated, they provide several benefits in scientific research, such as reproducibility, transparency, and collaboration. Version control allows different deductive paths to be explored and merged, facilitating collaboration among researchers with different expertise. Code versioning systems like GIT and SVN play a critical role in software development, enabling collaborative work, code reuse, and traceability. There also exist some dedicated solutions for versioning data such as DVC, DagsHub, Delta Lake, Dolt, Qri, Weights and Biases, Git LFS, Comet and LakeFS.

These different approaches are not suitable for our case study, since we want to work with concurrent versions and scenarios [2]. To answer even simple queries like **Q1**, current solutions require querying multiple database instances or checking out multiple version commits, which limits query response times.

2.2 Database versioning

A recent interesting solution called DoltHub, an online platform and hosting service provides version control for databases. This technology allows users to create, manage, and collaborate on databases using Git-style workflows and supports branching and merging, enabling teams to work on different features or versions of a database. However, cross-version querying of RDF data with such a solution is a challenge. Native SPARQL (SPARQL Protocol and RDF Query Language) queries are not inherently version-aware. Another solution called QuitStore [1] is an RDF data versioning system that addresses the need for efficient data retrieval across different versions. By implementing an RDF-based approach, QuitStore allows users to track changes and revisions to their semantic data over time. However, these versioning systems provide little support for cross-version queries. Indeed, they can either query metadata on multiple versions or query data on a single version.

Temporal databases, also known as historicized databases, are specialized databases that are designed to capture and store historical snapshots of data over a while. Some advanced temporal databases allow the analysis and querying of data at different points in time from two perspectives: how the data appeared in the real world and how it evolved within the database. However, by their nature, such databases are limited to a linear history and cannot be directly used to store a dataset with a branching history.

3 CONTRIBUTIONS PERSPECTIVES

Our motivation is to find a method for retrieving knowledge from a set of urban data versions stored in RDF format[4]. Resource Description Framework (RDF) offers a flexible and standardized format for representing the state of the city. RDF's graph-based structure allows the integration of diverse data sources, enabling a comprehensive view of the city's attributes and relationships. By versioning the city dataset, we can systematically track and record changes, modifications, and additions over time. This comprehensive version of history provides a foundation for analyzing the city's evolution, identifying trends, and extracting valuable knowledge. For example, if we have an urban dataset, we can identify the following problem: *How to analyze a set of versions of a city to produce additional knowledge?*

From a semantics point of view, a versioned graph can be assimilated to a collection of graphs, that is one graph for each version. Together with the `GRAPH` statement in SPARQL, this provides a way to query multiple versions at the same time. For example, the accessibility status of a given metro (**Q1**) or the height of the building (**Q2**). However separately storing each version would cost too much space and would probably lead to unefficient query processing.

Borrowing from historicized databases, one can associate version metadata to RDF triples. However, while a tuple in historicized database can be associated with a validity time interval, the branching nature of versioning history requires a different representation. Using provenance techniques [3] , this information could then be used at the query engine level to compute partial answers for several versions at once. We aim at implementing such a query engine and compare its efficiency with solutions using existing approaches with version checkout. We also aim at comparing the efficiency of different representations of version metadata associated to triples (for example representing the set of versions in extension or by a set of version intervals).

Note that this representation is independent from version metadata, we can thus reuse representation of version graph such as in [1] to trace the origin and lineage of data, for example to reference the code used to produce the data. It helps to understand its authenticity and assess its trustworthiness.

ACKNOWLEDGEMENTS

This work *Knowledge Hub for Evolving Urban Cities* is supported and funded by the IADoc@UDL (Université de Lyon, Université de Lyon 1) and LIRIS UMR 5205. We also acknowledge the BD team and the Virtual City Project[1] members for their invaluable advice and assistance.

REFERENCES

[1] Natanael Arndt. 2020. *Distributed Collaboration on Versioned Decentralized RDF Knowledge Bases.* Ph. D. Dissertation. Universität Leipzig. https://doi.org/10.33968/9783966270205-00

[2] John Samuel, Sylvie Servigne, and Gilles Gesquière. 2020. Representation of concurrent points of view of urban changes for city models. *Journal of Geographical Systems* (Feb. 2020). https://doi.org/10.1007/s10109-020-00319-1

[3] Leslie F. Sikos and Dean Philp. 2020. Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs. *Data Science and Engineering* 5, 3 (Sept. 2020), 293–316.

[4] Diego Vinasco-Alvarez, John Samuel, Sylvie Servigne, and Gilles Gesquière. 2021. Towards a semantic web representation from a 3D geospatial urban data model. In *SAGEO 2021, 16ème Conférence Internationale de la Géomatique, de l'Analyse Spatiale et des Sciences de l'Information Géographique.* 227–238. https://hal.science/hal-03240567/file/SAGEO_2021.pdf

Received June 7, 2023; revised August 21, 2023

[1] https://projet.liris.cnrs.fr/vcity/

Un graphe de données pour mesurer les Objectifs de Développement Durable et comprendre l'héritage des événements sportifs sur les villes

Wissal Benjira*
DVRC, Pôle Léonard de Vinci
LASTIG, Univ. Gustave Eiffel, IGN
Paris La Défense, France
wissal.benjira@{devinci.fr,ign.fr}

Faten Atigui
CEDRIC, CNAM
Paris, France
faten.atigui@lecnam.net

Bénédicte Bucher
LASTIG, Univ. Gustave Eiffel, IGN
Saint Mandé, France
benedicte.bucher@ign.fr

Malika Grim-Yefsah
LASTIG, Univ. Gustave Eiffel, ENSG
Marne la Vallée, France
malika.grim-yefsah@ensg.eu

Nicolas Travers
DVRC, Pôle Léonard de Vinci
Paris La Défense, France
nicolas.travers@devinci.fr

ABSTRACT

La disponibilité croissante de données joue un rôle majeur dans le suivi du développement durable des villes. Produire des indicateurs mesurant les Objectifs de Développement Durable (ODD) est un enjeu sociétal, nécessitant des sources hétérogènes, pour établir des comparaisons spatio-temporelles. Notre travail se focalise particulièrement sur l'étude de l'impact des événements sportifs sur les indicateurs de Développement Durable en mettant en évidence les enjeux managériaux liés à l'organisation de tels événements. Nous proposons une approche reposant sur un framework simplifié pour évaluer l'héritage laissé par ces événements, en utilisant un lac sémantique représenté par une base de données graphe stockée dans Neo4j. Cette approche est illustrée par une étude de cas basée sur des données variées, permettant de comparer l'état du territoire hôte avant, pendant et après la Coupe d'Europe de Football 2016 et d'évaluer l'évolution d'indicateurs dans différentes zones géographiques.

ACM Reference Format:
Wissal Benjira, Faten Atigui, Bénédicte Bucher, Malika Grim-Yefsah, and Nicolas Travers. . Un graphe de données pour mesurer les Objectifs de Développement Durable et comprendre l'héritage des événements sportifs sur les villes. In *Proceedings of 39ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA'23).*

1 CONTEXTE GÉNÉRAL

La disponibilité croissante de données couvrant des aspects variés de la réalité est une opportunité pour mieux observer et comprendre cette réalité dans sa complexité. Dans le domaine du développement durable urbain, ces données jouent un rôle crucial pour évaluer et mesurer l'impact des actions entreprises dans les villes [1, 2, 6]. Les Objectifs de Développement Durable (ODD) définis et reconnus par les pays membres des Nations Unies, mettent à disposition des indicateurs destinés à servir de fondement principal pour suivre les progrès vers la réalisation de ces ODD [8].

*Auteur correspondant

BDA'23, Octobre 2023, Montpellier, France

L'émergence d'événements sportifs, tels que les Jeux Olympiques, suscite un intérêt particulier en raison de leur influence significative sur les territoires hôtes. Ces évènements sportifs sont des phénomènes spatio-temporels qui affectent structurellement, économiquement et socialement un territoire accueillant ces évènements, générant ainsi un « héritage » [4, 5, 9]. L'étude de cet héritage repose sur la mesure comparative des différents impacts (ex. sociétaux, structurels, environnementaux, pratiques sportives, etc.) de ces événements à la fois sur une zone géographique et sur leurs évolutions au cours du temps.

L'étude de cet héritage sur les territoires nécessite différentes sources de données. Toutefois, la collecte des données peut s'avérer particulièrement complexe en raison de la diversité des sources, de la variété des formats, et finalement du manque de relations explicites entre les différentes données. L'examen de la littérature démontre une absence de cadre unificateur permettant de mobiliser des données pourvues d'hétérogénéités sémantiques.

Le sujet de la thèse s'intéresse plus précisément à la structuration de données et de métadonnées afin d'effectuer des analyses critiques et des comparaisons relatives à l'impact d'évènements sur les espaces urbains. Ainsi, l'objectif principal est de fournir un framework de croisement de données pour élaborer et exploiter des indicateurs associés aux ODD. Ce sujet prend tout son intérêt pour les collectivités territoriales, pour des porteurs de projets numériques autour des pratiques sportives et pour les sponsors de grands évènements.

2 MODÉLISATION PROPOSÉE

Dans la littérature, différentes approches de modélisation de schémas sont adaptées aux bases orientées documents ou colonnes, mais peu sur des bases de données graphes géolocalisées [7]. Ainsi, nous proposons une méthode de conception de lacs de données basée sur des bases de données graphes liés au contexte géomatique, i.e. intégrant des données géographiques, favorisant l'extraction et l'exploitation d'indicateurs ODD.

L'approche proposée repose sur un framework qui simplifie et facilite les différentes étapes d'évaluation de l'héritage en matière de développement durable; à savoir l'identification, le traitement et l'interprétation des données croisées pour obtenir des résultats plus

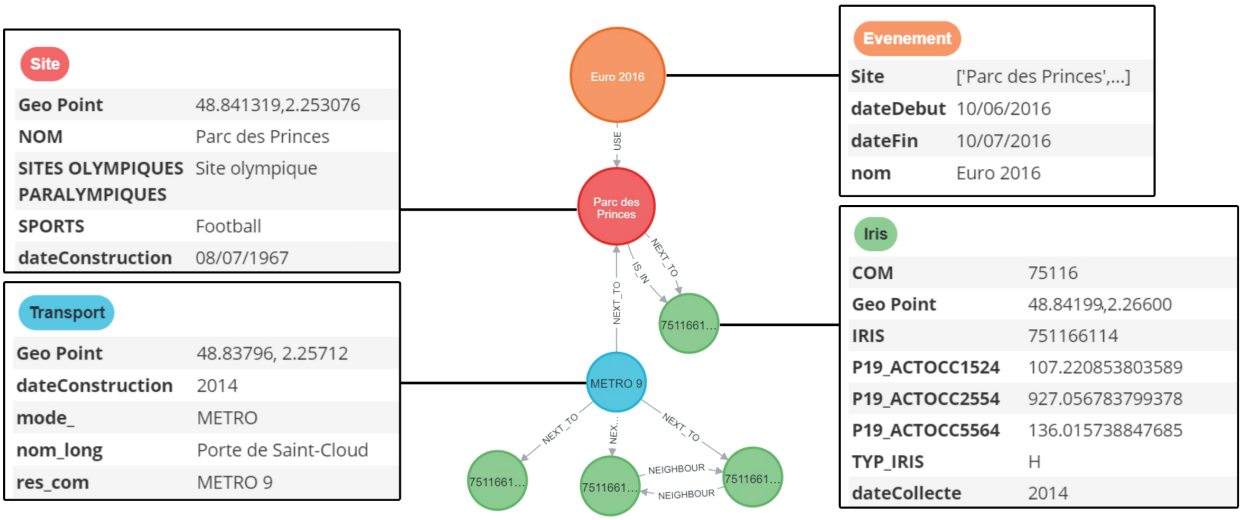


Figure 1: Sous-graphe Neo4j autour de l'événement sportif de la Coupe d'Europe de Football 2016

significatifs. Le lac sémantique obtenu est représenté et structuré sous forme de base de données graphe facilitant le rapprochement des données avec un schéma de graphe de données unifié, des règles d'enrichissement du graph et des opérations sur le graph pour produire des mesures et leur composition pour définir les indicateurs.

3 CAS D'ÉTUDE

Pour implémenter l'approche sémantique, nous développons une méthodologie de conception de base de données graphe guidée par les cas d'usages. Le prototype permet de valider et évaluer l'approche proposée. Nous avons choisi de considérer l'évaluation de l'impact de l'organisation d'événements sportifs sur l'indicateur du taux de chômage (commun à plusieurs ODD) en utilisant les différentes sources de données :

- Recensement de l'activité professionnelle par l'INSEE. La donnée est disponible à l'échelle des IRIS[1] et renseigne sur les caractéristiques des actifs, ex. le sexe, l'âge ou la catégorie socio-professionnelle.
- Géolocalisations des gares et stations par SmartIDF. Il s'agit d'un recensement des stations de transports en communs.
- Sites sportifs et installations sportives par SmartIDF. Il s'agit d'un recensement des installations sportives, ex. stades de football.

Pour illustrer, prenons l'événement « Coupe d'Europe de Football 2016 ». Nous avons pu établir le sous-graphe de la figure 1 offrant une vue globale des interconnexions entre les différentes entités.

Ce modèle de graphe facilite l'établissement d'indicateurs comparables dans le temps et dans l'espace, permettant ainsi des comparaisons du taux de chômage avant, pendant et après l'événement: l'héritage. Afin de manipuler le graphe et les indicateurs, le graphe est stocké dans Neo4j. Des requêtes Cypher interrogent aussi bien l'évolution temporelle du graphe que la géolocalisation du lieu. Cette géolocalisation est représentée à partir de sa description [3]. Par ailleurs, du fait que les indicateurs ODD reposent sur la composition de plusieurs mesures, notre Framework combine différentes requêtes pour calculer ces indicateurs de plus haut niveau.

Ainsi, l'évolution du graphe offre également la possibilité d'analyser ces indicateurs et d'observer les variations dans différentes zones géographiques. Les tendances démographiques, telles que les mouvements de population par sexe et par tranche d'âge, peuvent être analysées. De plus, l'analyse géospatiale permet d'identifier les zones où l'impact de l'événement est plus significatif. Les résultats obtenus peuvent être rapprochés à des cibles ODD et ainsi à des mesures d'héritages, permettant ainsi d'évaluer l'impact de l'événement par rapport à ces objectifs globaux de durabilité.

[1]IRIS - Îlots Regroupés pour l'Information Statistique

REFERENCES

[1] K. Anderson, B. Ryan, W. Sonntag, A. Kavvada, and L. Friedl. 2017. Earth observation in service of the 2030 Agenda for Sustainable Development. *Geo-spatial Information Science* 20 (2017).
[2] L. Ballerini and S. Bergh. 2021. Using citizen science data to monitor the Sustainable Development Goals: A bottom-up analysis. Sustain. *Sustain Sci* 16 (2021).
[3] H. Chen, M. Vasardani, S. Winter, and M. Tomko. 2018. A Graph Database Model for Knowledge Extracted from Place Descriptions. *ISPRS International Journal of Geo-Information* 7 (2018).
[4] M. Grim-Yefsah and B. Bucher. 2021. Towards Improving Knowledge Capitalization System for Sport Events Legacy. *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019) - KMIS* (2021).
[5] M. Harada. 2005. Hosting a mega-sports event and its impact upon city development challenges of the city of Osaka. Comparing Sports Policy, Sports Investment and Regional Development Initiatives in the Hosting of Sports Events in East Asia and Europe. (2005).
[6] K. Kurowska, R. Marks-Bielska, S. Bielski, A. Aleknavičius, and C. Kowalczyk. 2021. Geographic Information Systems and the Sustainable Development of Rural Areas. *Land* 10 (2021).
[7] J. Mali, S. Ahvar, F. Atigui, A. Azough, and N. Travers. 2022. A Global Model-Driven Denormalization. *Lecture Notes in Business Information Processing* 446 (2022).
[8] United Nations. 2015. Transforming our World: The 2030 Agenda for Sustainable Development. (2015).
[9] H. Preuss. 2019. Event legacy framework and measurement, International Journal Sport Policy Politics. *International Journal of Sport Policy and Politics* 11, 1 (2019), 103–118.

Apprentissage Distribué à Grande Echelle des Embeddings d'Ontologies avec OWL2Vec*

Yuhe Bai
yuhe.bai@lip6.fr
Sorbonne University
Paris, France

ABSTRACT

OWL2Vec*[1] est un algorithme conçu pour apprendre les embeddings d'ontologies en se basant sur des marches aléatoires afin de générer des séquences de mots, et en utilisant Word2Vec[7] pour calculer les embeddings. Bien que cet algorithme soit capable de produire des embeddings de haute qualité pour des tâches ultérieures, il présente actuellement une limitation en termes de scalabilité, car il traite de grandes ontologies sur une seule machine. Dans cette étude, nous visons à surmonter cette limitation en distribuant l'algorithme. Pour ce faire, nous partitionnons le graphe de connaissances, effectuons des marches aléatoires sur les sous-graphes, calculons les embeddings localement, puis réconcilions les différents espaces d'embeddings.

KEYWORDS

Algorithme distribué, Apprentissage distribué, Ontologies, Partitionnement de graphes, Marches aléatoires

1 INTRODUCTION

La représentation des graphes de connaissances ou *knowledge graphs* (KG) sous forme d'embeddings a suscité un grand intérêt au cours des dernières années [12]. Ces embeddings cherchent à représenter les éléments des KG, tels que les entités et les relations, dans un espace vectoriel qui préserve la structure du graphe. Un certain nombre d'algorithmes d'embeddings pour les KG ont été proposés et se sont avérés performants dans diverses tâches de prédiction.

Cependant, un grand nombre de ces algorithmes se concentrent sur les triplets de faits et sont insuffisants pour traiter les ontologies OWL ou les schémas ontologiques exprimés en OWL. Ces ontologies sont plus complexes que les simples structures de graphes, intégrant des opérateurs logiques tels que la disjonction de classe, la quantification existentielle et universelle et des métadonnées qui comprennent des synonymes, des définitions et des annotations relatives à une classe.

OWL2Vec*[1] est l'une des premières tentatives pour intégrer la sémantique des ontologies OWL, y compris la structure des graphes, les littéraux et les opérateurs logiques. Sa conception générale a été conçue pour gérer de manière flexible diverses sémantiques OWL.

Bien que cet algorithme d'embedding d'ontologies soit capable de produire des embeddings de haute qualité pour diverses tâches prédictives, il manque de capacité pour la formation à grande échelle sur plusieurs machines. Nous proposons donc un cadre d'apprentissage distribué.

Dans le contexte des algorithmes distribués, il est crucial de maintenir la qualité des embeddings générés. De plus, la communication entre machines peut être un goulot d'étranglement significatif. En résumé, la distribution d'OWL2Vec* présente des défis liés au partitionnement des graphes, aux marches aléatoires sur les sous-graphes, à la gestion des coûts de communication, et à la garantie de haute qualité des embeddings.

2 TRAVAUX CONNEXES

De nombreuses méthodes ont été proposées pour apprendre les embeddings de graphes de connaissances, comme TransE[4], DistMult[5], et ComplEx[11]. Concernant les embeddings d'ontologies, des algorithmes comme Onto2Vec[8], OPA2Vec[9], et OWL2Vec*[1] ont été développés, ce dernier offrant une intégration robuste de la sémantique des ontologies OWL. Des frameworks distribués[3] ont également été explorés pour l'apprentissage d'embeddings sur plusieurs machines.

Cependant, les systèmes existants ont des limites pour traiter les ontologies avec des expressions OWL et ont des coûts de communication élevés. Dans notre travail, nous visons à améliorer le passage à l'échelle d'OWL2Vec*[1] avec une version distribuée de l'algorithme, introduisant un processus de partitionnement du graphe et de réconciliation des espaces d'embeddings, tout en réduisant les coûts de communication.

3 MÉTHODOLOGIE

Nous nous appuyons sur la méthode proposée par [3] pour distribuer le calcul d'embeddings de graphes. Cette méthode consiste à choisir un petit nombre de noeuds appelés points de repère ou *landmarks* et qui serviront par la suite à réconcilier les embeddings. Nous adaptons cette méthode pour distribuer le calcul des embeddings de KG définis dans OWL2Vec*[1].

Etape 1: Partitionnement du graphe de connaissances
La première étape consiste à partitionner le graphe de connaissances. Pour un premier temps, nous utilisons l'algorithme de partitionnement classique graph-cut METIS[6] pour cette tâche. Les nœuds partagés sont utilisés comme *landmarks* et aideront à l'étape de réconciliation. Nous explorons actuellement d'autres techniques de partitionnement que METIS qui pourraient s'avérer bénéfiques et plus adaptées aux graphes de connaissances.

Etape 2: Extraction de marches aléatoires et entraînement de Word2vec
Après avoir partitionné le graphe de connaissances, nous extrayons des marches aléatoires de chaque sous-graphe et entraînons Word2vec[7] localement sur chaque partition, y compris les *landmarks*. Ainsi,

chaque partition a sa propre représentation d'un espace latent capturant les informations spécifiques à cette partition.

Nous explorons actuellement d'autres méthodes que Word2Vec pour le calcul des embeddings. Les modèles de langue pré-entrainés tels que BERT[2] pourraient être capable de représenter, de manière complémentaire à OWL, la sémantique d'un graphe de connaissances.

Etape 3: Calcul de la matrice de transformation et réconciliation

Les nœuds partagés entre deux partitions auront des embeddings distincts dans leurs espaces latents respectifs. Nous réutilisons la stratégie de réconciliation proposée dans [3]. Cette méthode détermine une "rotation" permettant d'associer un premier espace d'embeddings à un autre : Une matrice de transformation W est calculée à partir des embeddings des landmarks communs aux deux espaces en utilisant une décomposition en valeurs singulières (SVD). Ensuite les embeddings "réconciliés" sont obtenus en multipliant par W les embeddings du premier espace. La Figure 1 illustre les étapes de notre approche.

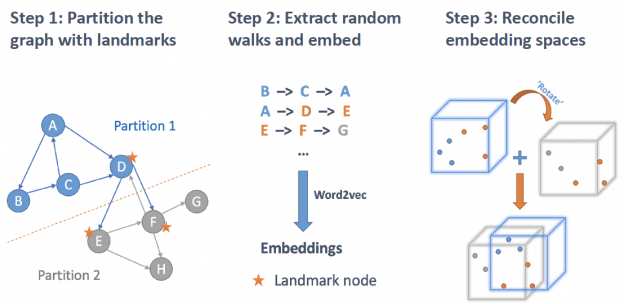


Figure 1: Principales étapes de notre méthode

4 RÉSULTATS EXPÉRIMENTAUX

Le tableau 1 présente les résultats obtenus pour deux partitions, sans et avec réconciliation, en utilisant le jeu de données FoodOn pour la prédiction de subsumption de classe (Prédiction d'une classe parente d'une classe). On peut observer que la réconciliation améliore d'environ 30% la qualité de la prédiction par rapport à la méthode sans réconciliation, tout en maintenant un apprentissage entièrement localisé et un coût de communication faible.

Table 1: Résultat sur la tâche de class subsumtion prédiction sur FoodOn

	MRR	Hits@1	Hits@5	Hits@10
No Reconciliation	12.8	7.0	17.3	23.0
With Reconciliation	**16.4**	**10.1**	**22.8**	**28.6**

Nous poursuivons la validation de notre approche en considérant des jeux de données plus grands, en explorant d'autres tâches comme la prédiction de liens et le typage d'entités, et en étudiant l'impact du partitionnement sur la méthode de réconciliation et la qualité des résultats.

5 CONCLUSION

Dans ce travail, nous avons apporté plusieurs contributions :

(1) nous avons distribué l'algorithme OWL2Vec*[1] avec succès en utilisant l'algorithme de partitionnement de graphes METIS[6], et réconcilié les embeddings avec l'aide de la décomposition en valeurs singulières (SVD).
(2) La réconciliation a montré son efficacité en produisant des embeddings de haute qualité et en réduisant les coûts de communication.

Pour les travaux futurs, nous envisageons d'optimiser notre sélection de *landmarks* et d'évaluer notre méthode sur d'autres graphes de connaissances à grande échelle comme YAGO[10], pour tester la polyvalence et l'efficacité de notre méthode.

6 RECONNAISSANCE

Je tiens à remercier le Centre d'Intelligence Artificielle de la Sorbonne (SCAI) pour le financement partiel de cette recherche par le biais d'une bourse doctorale. Je tiens également à remercier mon directeur de thèse, Hubert Naacke, ainsi que ma co-encadrante, Camelia Constantin, pour leurs précieux conseils tout au long de mes travaux de recherche.

REFERENCES

[1] Jiaoyan Chen, Pan Hu, Ernesto Jimenez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, and Ian Horrocks. 2021. Owl2vec*: Embedding of owl ontologies. *Machine Learning* 110, 7 (2021), 1813–1845.
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]
[3] Chi Thang Duong, Trung Dung Hoang, Hongzhi Yin, Matthias Weidlich, Quoc Viet Hung Nguyen, and Karl Aberer. 2021. Scalable robust graph embedding with Spark. *Proceedings of the VLDB Endowment* 15, 4 (2021), 914–922.
[4] Bordes et al. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013).
[5] Yang et al. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575* (2014).
[6] George Karypis and Vipin Kumar. 1997. METIS: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices. (1997).
[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL]
[8] Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. 2018. Onto2vec: joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics* 34, 13 (2018), i52–i60.
[9] Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. 2019. OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics* 35, 12 (2019), 2133–2140.
[10] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*. 697–706.
[11] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *Proceedings of The 33rd International Conference on Machine Learning*. 2071–2080.
[12] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering* 29, 12 (2017), 2724–2743.

Entropy Maximisation For Diverse Recommendations

Jonathan Colin
jonathan.colin@universite-paris-saclay.fr
Paris Saclay University
Gif-sur-Yvette, France

ABSTRACT

The popularity of online social networks and the interactions they allow has brought great benefits in terms of ease of communication between humans. On the other hand, it has been found that they may also be a disruptive force for society, notably through the spread of fake news or the creation of filter bubbles. Indeed, in order to maximize user engagement, recommender systems tend to overrate trendy content or content that is similar to a user's established taste; this overshadows any form of new, niche, or diverse content. This is a critical issue as these processes end up increasing opinion polarization. In this paper, in order to alleviate this issue, we propose a new se- mantic for diverse link recommendations in online social networks based on increasing the degree of information entropy. We perform a comparative analysis of our proposed semantic to baseline tra- ditional semantics for recommendation and find that traditional recommender systems tend to reduce information entropy. At the same time, experimental results show that the proposed semantic manages to promote diverse content.

KEYWORDS

social networks, echo chambers, diversity, optimization, recommender systems

8 Prix BDA 2023

8.1 Prix des articles de recherche

BDA a la particularité de proposer deux catégories d'articles : les articles originaux non publiés et les articles publiés récemment dans une conférence internationale de renom. Cette dernière catégorie permet de diffuser largement les travaux faisant la renommée internationale de notre communauté nationale en gestion de données.

Lauréats du prix des articles de recherche

Scalable Reasoning on Document Stores via Instance-Aware Query Rewriting, *Olivier Rodriguez, Federico Ulliana, Marie-Laure Mugnier*

GPC : A Pattern Calculus for Property Graphs, *Nadime Francis, Amélie DR Gheerbrant, Paolo Guagliardo, Leonid Libkin, Victor Marsault, Wim Martens, Filip Murlak, Liat Peterfreund, Alexandra Rogova, Domagoj Vrgoc*

8.2 Prix des démonstrations

Lauréat du prix des démonstrations

Headwork : Powering the Crowd with Tuple Artifacts, *David GROSS AMBLARD, Constance Thierry*

8.3 Prix des thèses en gestion de données

Lauréats du prix des thèses

Prix de thèse : Quentin Manière pour sa thèse intitulée
« *Counting Queries in Ontology-Based Data Access* ».
encadrants : Meghyn Bienvenu et Michaël Thomazo

Accessit au Prix de Thèse : Daniel Rosendo pour sa thèse intitulée
« *Methodologies for Reproducible Analysis of Workflows on the Edge-to-Cloud Continuum* »
encadrants : Gabriel Antoniu, Alexandru Costan et Patrick Valduriez