EURA NOVA

FACULTIES OF SCIENCE AND ENGINEERING

Master Thesis Offers 2022-2023

For the latest information, please visit https://research.euranova.eu.
20th November 2021

CONTENT

Euranova

Introduction

Euranova is a data-driven Belgian company founded in September 2008 and located in Brussels, Marseille, and Tunis. Our mission is simple: bring life to our customers' great ideas by offering best-in-class services in data science, software engineering, and data architecture. To do so, we invest significantly in in-house expertise and state-of-the-art knowledge. In line with this course of action, we offer academic programs in collaboration with universities. These offers include boot camps, master theses topics, research internships, and PhDs topics. See below for details.

Our Master Theses Offers

This document presents master theses and graduation project topics supervised by our research & development department. Each project is an opportunity to be actively involved in the development of solutions to address tomorrow's challenges in ICTs and implement them today. The students will work with a dedicated international team of engineers with diverse expertise in machine learning, graph theory, artificial intelligence, high-performance computing, etc. They will keep Eura Nova informed of the project advancement and share their ideas and challenges using the in-house knowledge management tool. We value continuous learning and teamwork. We love to have a good time together. For more information on our R&D activities, please visit our website at https://research.euranova.eu.

How To Apply

When you have gone through our master thesis offers, pick your favourite. Draft a short text stating why you find it interesting and what you would do about it. Send us this statement, along with your CV at career@euranova.eu. If you are interested in working on a topic that is not in our range of offers, we would be delighted to hear your proposition and invite you to get in touch as well.

Automatic Generation of Natural Language Privacy Policies from Knowledge Bases

Context

Privacy policies are the documents describing the processing that a Data Controller is conducting with the personal data of a Data Subject. According to the General Data Protection Regulation (GDPR), such documents as privacy policies should state clearly and unambiguously what kind of actions are performed, with which category of personal data, for which purpose and by whom. Additionally, the processing should be justified by a legal basis and protected by specific technical and organisational measures.

The objective of these documents is to be informative, for legal reasons, yet simple enough to be understood by the users. And creativity in writing privacy policies does not always bring the best outcome. While becoming more “fun” for the users to read, they may also be difficult to analyze from a legal standpoint. The analysis of contracts and policies is an important task for a Data Protection Officer (DPO).

This Master Thesis offers to facilitate both types of privacy policy readers - regular Data Subjects and DPOs/lawyers. By using a machine-readable representation of policies, the main objective is to generate human-readable privacy policies which are clear and simple. Our model - Semantic dAta priVacy modEl (SAVE) [1] - consists of rules (permissions, prohibitions and obligations) describing personal data processing. Currently, the model is used to extract the rules from the natural language text of privacy policies and represent them as a knowledge base in RDF format. This Master Thesis is supposed to answer the opposite question: “How to generate natural language policy using a knowledge base?”. The final goal is to generate complete and coherent textual documents that would be 1) understandable by humans and 2) acceptable as official privacy policies. The latest transformer-based models can be used to generate natural language textual documents [2], including text generation from knowledge bases [3,4]. Adapting such models or taking inspiration from them may provide us with an innovative solution for privacy policy generation.

Business opportunities

Automatic generation of privacy policies can significantly facilitate the tasks of a DPO, making the creation and updates of privacy policies easy and straightforward. Additionally, the clear, easily understandable text generated by the model will make the readers’ lives easier. The resulting model can be wrapped in a library or a tool available for use by DPOs and companies.

Contribution

The objectives are the following:

- Exhaustive state-of-the-art study on Natural Language Generation (NLG): traditional methods, the latest models using transformers
- Exhaustive state-of-the-art study on NL generation from ontologies and knowledge bases
- Study of the current techniques of privacy policy writing/generation.
- Designing and implementing a novel NLG model for privacy policy generation, by applying or adapting state-of-the-art NLP models.
- Evaluating the model: defining the metrics, running the experiments, analysing the results, drawing conclusions.

Technologies/expertise to develop

- Python
- Natural Language Generation with Transformer models: BERT, GPT, T5, etc.
- Semantic Web: RDF, OWL, Knowledge Bases

In practice

Belgium, Academic year 2022-2023

References

- [1] Krasnashchok, Katsiaryna, Majd Mustapha, Anas Al Bassit, and Sabri Skhiri. "Towards Privacy Policy Conceptual Modeling." In *International Conference on Conceptual Modeling*, pp. 429-438. Springer, Cham, 2020.
- [2] Topal, M. Onat, Anil Bas, and Imke van Heerden. "Exploring transformers in natural language generation: GPT, BERT, and XLNet." *arXiv preprint arXiv:2102.08036* (2021).
- [3] Koncel-Kedziorski, Rik, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. "Text generation from knowledge graphs with graph transformers." *arXiv preprint arXiv:1904.02342* (2019).
- [4] Yu, Wenhao, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. "A survey of knowledge-enhanced text generation." *arXiv preprint arXiv:2010.04389* (2020).

Automating Data Quality Rules From Knowledge Graph

Context

Knowledge and semantic graphs have been used to model the enterprise data model to better manage data lake or data warehouse [1]. Recently, these graphs have also been used for on-demand data integration [2] or even ETLs for EDWH [3]. And as this knowledge graph represents the underlying data structure, it has also been used for data integration in streaming [4].

Having a knowledge graph for representing data opens the doors to new applications. One of them is the automation of data quality rules. For example, in the insurance sector, you need a customer to always be associated with the right account. It is a constraint that requires technical data quality. Existing constraint languages such as SHACL [5], used by Euranova for GDPR compliance checking, could represent the constraints on data and then infer the data quality KPI.

Business opportunities

Data quality is a very hot topic, but it is also increasingly complicated as the data volume never stops growing. It is becoming almost impossible to cover the entire data set with good quality KPI. Consequently, data quality automation is one of the biggest challenges in data management.

Contribution

The student will first study the state of the art of knowledge graph in data management, then focus on data quality management. The first contribution will be to propose a method, model or algorithm to automate the SHACL constraint definitions from the knowledge graph. The second contribution will focus on how to evaluate the quality KPI using the SHACL language and first-order logic.

Technologies/expertise to develop

- Knowledge graph
- Data quality
- SHACL constraint language
- First order logic
- Data management

In practice

Belgium, Academic year 2022-2023.

References

[1] Shumet Tadesse and Cristina Gomez and Oscar Romero and Katja Hose and Kashif Rabbani, . "ARDI: Automatic Generation of RDFS Models from Heterogeneous Data Sources." . In 23rd IEEE International Enterprise Distributed Object Computing Conference, EDOC 2019, Paris, France, October 28-31, 2019 (pp. 190–196). IEEE, 2019.
[2] S. Nadal, A. Abello, O. Romero, S. Vansummeren and P. Vassiliadis, "Graph-driven Federated Data Management," in IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2021.3077044.
[3] Rudra Pratap Deb Nath and al., High-Level ETL for Semantic Data Warehouses. June 2020.

[4] Belcao M., Falzone E., Bionda E., Valle E.D. (2021) Chimera: A Bridge Between Big Data Analytics and Semantic Technologies. In: Hotho A. et al. (eds) The Semantic Web – ISWC 2021. ISWC 2021. Lecture Notes in Computer Science, vol 12922. Springer
[5] https://www.w3.org/TR/shacl/

Automatic Knowledge Graph Construction from Richly Formatted Documents

Context

In the context of a digital economy, large amounts of data regarding different aspects of the parties' commercial relationships flow through contractual networks formed between companies.

Privacy regulations often create high standards for the data protection of companies and citizens. Usually, this regulatory framework is difficult to enforce, given the difficulties related to monitoring data. Particularly in situations where the processing activities of the companies supplying data involve a significant number of subcontractors, generating a long supply chain, it is a challenge to ensure compliance with data protection regulations.

Recent regulations, such as the General Data Protection Regulation and California Consumer Privacy Act, seek to empower consumers by granting additional privacy rights and limiting the number of processing activities that can be carried out without explicit consent. Consequently, current research on data protection and privacy is predominantly data-subject-centric and focuses mainly on B2C relationships. Instead, this project studies data processing activities up and downstream of the supply chain. It contends that individuals' right to privacy may lack effectiveness if it is not possible to hold processors accountable for the data flows in the supply chain.

Recent studies on privacy compliance, automation, and machine learning considered only privacy policies and have not yet investigated companies' internal documents to document data processing activities. We aim to fill this gap by including in our study data protection artefacts generated throughout the data supply chain, such as the Data Protection Addendum and emails.

Business opportunities

Using AI methods to augment data protection officers (DPOs) while conducting data protection compliance tasks can significantly increase their productivity, thus, focusing on more complex tasks. Consequently, this will positively impact the business and decrease the risk of fines, enforce regulations' compliance, and support the adoption of privacy by design.

Contribution

The objectives are the following:

- Exhaustive state-of-the-art study on Knowledge base construction, Knowledge graphs, and Information extraction from richly formatted documents.
- Design of the knowledge graph schema with help from the data protection experts. This schema will define the entities' classes, attributes, and relationships.
- Development of neural methods based on pretrained transformers to automatically populate the knowledge graph.
- Verification of the key requirement[1], established by the European Commission, to ensure that the AI methods output by this project is trustworthy.

Technologies/expertise to develop

- Knowledge base construction

[1] Human agency and oversight, Technical Robustness and safety, Privacy and data governance, Transparency, Diversity, non-discrimination and fairness, Societal and environmental well-being, Accountability.

- Knowledge graphs
- Information extraction from richly formatted documents
- Data programming

In practice

Belgium, Academic year 2022-2023.

References

[1] Bollacker, K. a. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. Proceedings of the 2008 ACM SIGMOD international conference on Management of data, (pp. 1247-1250).
[2] He, Q. a. (2016). Building the linkedin knowledge graph. Engineering. linkedin. Com.
[3] Kocayusufoglu, F. a. (2019). RiSER: Learning Better Representations for Richly Structured Emails. The World Wide Web Conference, (pp. 886-895}).
[4] Minervini, P. a. (2020). Differentiable Reasoning on Large Knowledge Bases and Natural Language.
[5] Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web, 489-508.
[6] Ratner, A. J. (2016). Data programming: Creating large training sets, quickly. Advances in neural information processing systems, (pp. 3567-3575).

How Privacy-Preserving Technologies In Neural Networks & Graph Neural Networks ?

Context

The EU has pushed privacy-preserving technologies [1] to foster innovation around data while preserving privacy. Different technologies have been emerging from this initiative. Synthetic data and encryption are the most well-known ones, especially FHE (Full Homomorphic Encryption) [2]. The basic idea of the FHE is that you can apply (linear) operations on encrypted data, and the result (encrypted) is the same as if the operation was applied on clear data and then encrypted.

But, this is only recently that we have seen the application of this method to non-linear models such as neural networks.

Business opportunities

PET (Privacy Enhancing Technologies) in data management opens new doors in terms of the data ecosystem. We would be able to share encrypted private data with partners without compromising the data. On the other hand, the partners will be able to analyse data and create predictive models without infringing the law.

Contribution

The objectives of this thesis are

1. To study the PET and their current maturity
2. To study the PET usage in machine learning and especially in neural networks as [3]
3. To study the PET usage in graph neural networks
4. To propose a new method for using PET in graph neural networks and benchmarking the result

Technologies/expertise to develop

- PET and encryption
- Privacy by design
- Graph neural networks

In practice

Belgium, Academic year 2022-2023.

References

[1] https://www.enisa.europa.eu/topics/data-protection/privacy-enhancing-technologies
[2] Hamlin, A., Schear, N., Shen, E., Varia, M., Yakoubov, S., & Yerukhimovich, A. (2016). Cryptography for Big Data Security. Cryptology ePrint Archive, Report 2016/012.
[3] Onoufriou and al. EDLaaS; Fully Homomorphic Encryption Over Neural Network Graphs. Published on arxiv 2021.