Semantic Graph Mining for Black-Box Optimisation

Proposition de projet de recherche doctoral

Campagne SCAI 2022

Thesis supervisors
Marie-Jeanne Lesot, LIP6, LFI team, axis: Artificial intelligence and data science
Carola Doerr, LIP6, RO team, axis: Theory and mathematics of computing
Affiliation
Lab: LIP6, UMR7606, SU, CNRS
Ecole Doctorale: ED130 - EDITE
Keywords: ontology, conceptual graphs, graph pattern mining, fusion and aggregation, black-box optimisation, algorithm selection, benchmarking, XAI

Abstract

The general aim of the thesis is to exploit expert knowledge regarding properties of optimisation algorithms and problems, represented in the formal frameworks of ontologies and conceptual graphs, and to develop tools to extract automatically underlying correlations: the objective is to allow understanding the reasons why an algorithm is more appropriate than others to solve a problem depending on its characterisation and possibly to offer new tools to configure optimisation algorithms.

To do so, the project will explore new methods for analysing conceptual graphs and in particular design dedicated frequent pattern mining algorithms. The output of such algorithms is also represented in the framework of conceptual graphs, which makes the results more legible and understandable for the end user.

The thesis is expected to contribute at the cross-roads of the domains of knowledge representation, pattern mining and black-box optimisation.

Context

With the rise of artificial intelligence, exploiting expert knowledge in order to explain or understand the observed results and performances is at high demand. To enable systematic studies of rich data collections, we need both a suitable formalism to represent the available knowledge and efficient tools to exploit them.

The thesis project aims at addressing these questions for the domain of black-box optimisation, in the case where knowledge is represented in the formalism of conceptual graphs and using tools in the form of frequent pattern mining.

Knowledge representation To store knowledge, ontologies constitute a currently very popular approach, as they enable to define the ideas and concepts manipulated in any specific domain and then reason on these concepts in order to infer new knowledge about them. Together with the associated reasoning engine, they are in particular useful for classification purposes: they allow one to get the class to which an object belongs, according to its characteristics.

Conceptual graphs (see e.g. Chein & Mugnier, 2008) constitute a richer knowledge representation framework, that explicitly distinguishes between ontological and factual knowledge: the former defines a vocabulary, in the form of a hierarchically structured terminology, used to build the latter, in the form of bi-partite graphs whose nodes and edges are labelled with elements from the vocabulary. Conceptual graphs offer the advantage of combining a logical semantic and efficient manipulation tools based on graph theory.

Frequent pattern mining Extracting knowledge from a conceptual graph or from a set of conceptual graphs can take the form of identifying frequent patterns, i.e. subgraphs that occur frequently and can thus be interpreted as relevant regularities. This relates to the domain of Frequent Subgraph Mining, for which numerous approaches have been proposed (see e.g. Jiang et al., 2013). Dedicated algorithms have been

proposed for the specific case of taxonomy labelled graphs (Inokuchi et al., 2000; Cakmak & Ozsoyoglu, 2008; Petermann et al., 2017), that exploit the particular characteristics of these types of graphs so as to improve both their efficiency and the relevance of the extracted patterns.

Recent works conducted in the LFI team focused further on the case of conceptual graphs: the cgSpan algorithm (Faci et al., 2021a) exploits the information about relation arity to decrease efficiently the number of explored pattern candidates, the relation signatures to prune redundant pattern candidate with little informativeness as well as inference rules to extend candidates faster.

Ontologies of optimisation algorithms and problems To develop our approaches and to show their benefits on real-world knowledge graphs, we will test them on the OPTImisation algorithm benchmarking ONtology (OPTION, Kostovska et al., 2021), available at the BioPortal.[1]

OPTION groups and annotates benchmarking data from black-box optimisation. More specifically, its current version hosts more than 200 algorithm performance data from the BBOB collection of the COCO framework (Hansen et al., 2020) and from various benchmark suits of the Nevergrad environment (Rapin & Teytaud, 2018). OPTION provides the vocabulary needed for semantic annotation of the core entities involved in the process of benchmarking black-box optimisation algorithms, such as algorithms, problems, and evaluation measures. OPTION also provides means for automated data integration, improved interoperability, powerful querying capabilities and reasoning, thereby enriching the value of the benchmark data.

A core advantage for our project is that OPTION builds a conceptual graph with a rich vocabulary, with deep concept and relation hierarchies. It therefore offers a broad number of use-cases on which we can develop and validate our approaches. To demonstrate the advantages of our algorithms, and to contribute to an improved use of benchmarking data in the analysis of black-box optimisation algorithms, we directly work with the developers of OPTION, among them the second supervisor of this PhD project.

Scientific Objectives

The key objective of the proposed thesis project, from the black-box optimisation point of view, is to investigate the potential of applying proper data-mining techniques to the OPTION ontology, with the goal to derive recommendations for algorithm selection (Kerschke et al., 2019), in an automated and ideally continuously extendable fashion.

To do so, the thesis will explore new methods for analysing conceptual graphs, for instance to deal with the fact that the input here takes the form of a unique huge graph. Indeed, existing approaches mainly consider the case where multiple graphs are given as inputs. Besides, the considered application calls for the design of incremental information extraction approaches, that are able to process an enriched conceptual graph without starting from scratch whenever the conceptual graph is updated.

Another direction of research will aim at defining formally the quality criteria for frequent conceptual subgraphs, that raise challenging questions due to the available taxonomies over the labels: an increase of the level of generality of the labels is correlated with an increase of the candidate frequency, but also with a decrease of its informativeness for the end user. Thus the global quality must take into account several components and trade-off thereof. Designing extraction algorithm to mine efficiently and directly patterns optimising such a combined quality criterion opens an additional direction to be explored.

In addition, in the considered application domain, rare subgraph patterns can also be of high relevance to an expert user, to draw his/her attention to specific optimisation algorithms configurations or specific characteristics of optimisation problems. Designing methods able to identify them, efficiently exploiting the properties of conceptual graphs, will be considered during the thesis.

Finally, especially because the OPTION includes nodes with numerical labels, the thesis will consider the question of representing and building fuzzy extensions of conceptual graphs (see e.g. Thomopoulos et al., 2003; Faci et al., 2021b), or more generally weighted extensions: including some imprecision in the node labels may be a relevant manner to generalise nodes and to define more relevant frequent patterns. Other semantics for the weights will be explored, e.g. to represent optional parts of patterns or certainty degrees. The challenges associated with this topic are both to determine whether existing weighted conceptual graphs appropriately capture the desired semantics and to design algorithms able to extract, or to build, such graphs automatically.

[1] https://bioportal.bioontology.org/ontologies/OPTION-ONTOLOGY

Adequacy to the Sorbonne Centre for Artificial Intelligence Initiative

The proposed topic involves AI in several ways, as both frequent pattern mining as well as black-box optimisation are core AI subjects. The automated selection and configuration of black-box algorithms is a heavily studied in so-called AutoML context (Hutter et al., 2019), to which we will contribute a conceptually different approach. The use of conceptual graphs, both for the input and output representation, contributes to the human-in-the-loop paradigm of XAI.

Supervision

This project continues an ongoing collaboration between the two involved teams at LIP6. While Marie-Jeanne has been working on knowledge representation and knowledge extraction through machine learning (see http://webia.lip6.fr/~lesot/publications.html), Carola brings expertise in black-box optimisation (see http://www-ia.lip6.fr/~doerr/ for CV and research activities).

PhD candidate profile

A Master's degree in a quantitative field such as Computer Science, Engineering, Statistics, Operations Research, Mathematics is required. We expect willingness to conduct empirical research as well as experience with the python programming language. Since the student will be working in an international research team, they must be proficient in written and spoken English. Knowledge of French is not required. International students are very welcome to apply.

References

Cakmak, A., & Ozsoyoglu, G. (2008). Taxonomy-superimposed graph mining. *Proc. of the 11th Int. Conf. on Extending database technology: Advances in database technology* (pp. 217–228).

Chein, M., & Mugnier, M.-L. (2008). *Graph-based knowledge representation: Computational foundations of conceptual graphs*. Springer.

Faci, A., Lesot, M.-J., & Laudy, C. (2021a). cgspan: Pattern mining in conceptual graphs. *Proc. of the Int. Conf. on Artificial Intelligence and Soft Computing* (pp. 149–158). Springer, LNCS12855.

Faci, A., Lesot, M.-J., & Laudy, C. (2021b). Fuzzy conceptual graphs: a comparative discussion. *Proc. of the Symposium Series on Computational Intelligence, Foundation of Computational Intelligence (SSCI-FOCI21)*.

Hansen, N., Auger, A., Ros, R., Mersmann, O., Tušar, T., & Brockhoff, D. (2020). COCO: A platform for comparing continuous optimizers in a black-box setting. *Optimization Methods and Software*, *36*, 114–144.

Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.). (2019). *Automated machine learning - methods, systems, challenges*. The Springer Series on Challenges in Machine Learning. Springer.

Inokuchi, A., Washio, T., & Motoda, H. (2000). An apriori-based algorithm for mining frequent substructures from graph data. *Proc. of the European Conf. on principles of data mining and knowledge discovery* (pp. 13–23).

Jiang, C., Coenen, F., & Zito, M. (2013). A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review*, *28*, 75–105.

Kerschke, P., Hoos, H. H., Neumann, F., & Trautmann, H. (2019). Automated algorithm selection: Survey and perspectives. *Evolutionary Computation*, *27*, 3–45.

Kostovska, A., Vermetten, D., Doerr, C., Dzeroski, S., Panov, P., & Eftimov, T. (2021). Option: optimization algorithm benchmarking ontology. *Proc. of the Genetic and Evolutionary Computation Conference Companion, GECCO Companion 2021* (pp. 239–240).

Petermann, A., Micale, G., Bergami, G., Pulvirenti, A., & Rahm, E. (2017). Mining and ranking of generalized multi-dimensional frequent subgraphs. *Proc. of the 12th Int. Conf. on Digital Information Management (ICDIM)* (pp. 236–245).

Rapin, J., & Teytaud, O. (2018). Nevergrad - A gradient-free optimization platform. https://GitHub.com/FacebookResearch/Nevergrad.

Thomopoulos, R., Buche, P., & Haemmerlé, O. (2003). Representation of weakly structured imprecise data for fuzzy querying. *Fuzzy Sets and Systems*, *140*, 111–128.