Patent Valuation with Graph Neural Networks

Representation Learning on Patent Knowledge Graphs for Value Prediction

Graduate

Moritz Bättig

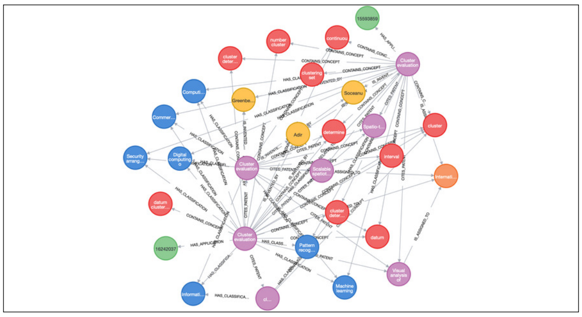
Definition of Task: Patents sit at the intersection of law, technology, and economics. They not only protect inventions but also influence investment, licensing, and litigation. This makes their valuation a key interest for companies, investors, and researchers. Machine learning has recently opened new approaches to this problem. Graph Neural Networks (GNNs) are well suited for learning from interconnected data. Viewing the patent system as a network — where nodes represent patents and edges capture citations, shared inventors, technology classes, or co-ownership — a GNN can learn from both a patent's content and its position in the network. This includes information such as citations, related patents by the inventor, technologies it connects, and the importance of its owner. This thesis examines whether building a knowledge graph and training a GNN can improve patent valuation predictions compared to traditional methods.

Approach: The thesis is divided into two main parts. First, patent data was collected from the US Patent and Trademark Office (USPTO) for two technical domains, quantum computing and machine learning, to build two knowledge graphs. Additional patent quality indicators from the OECD were included, and ground-truth valuations were taken from the market-based measure by Kogan et al. (KPSS, 2017). Technical ideas and concepts were extracted from the documents to group patents with shared concepts, such as "error correction" in quantum computing. The knowledge graph was then constructed in Neo4j with Patent, Inventor, Application, Concept, Classification, and Assignee nodes, linked through their relationships. Eleven patent value indices, proposed in prior research, were calculated from the graph, and categorical features were one-hot encoded while textual features were embedded into dense vectors.

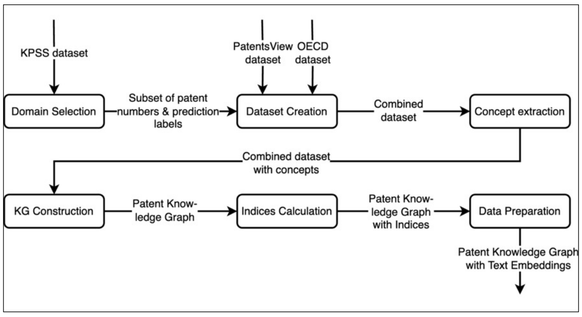
After constructing the graph, machine learning models were implemented to predict patent value. Baseline models — a Random Forest regressor and a simple feedforward neural network — were trained on a CSV export of patent node properties without graph topology. Next, FastRP and GraphSAGE were applied in Neo4j to generate node representations that incorporated both properties and neighborhood information. Those representations were then used in separate downstream prediction models. Finally, an end-to-end Graph Neural Network was trained in PyTorch Geometric using HGT, HAN, and SAGE architectures. Due to the underperformance of the graph-based methods in regression, additional binary classification models were trained using the same techniques.

Result: The results show that the graph-based approaches did not outperform the baseline models in any case. In patent valuation, more complex models such as GNNs do not necessarily lead to better results. The Random Forest model, trained directly on tabular features, outperformed all others, suggesting that node features alone contain strong predictive signals. Complex GNNs are also prone to over-smoothing or over-squashing neighboring node properties, which can cause information loss and poor performance. Predicting exact patent values proved difficult, with weak regression results across all models. Switching to a binary classification task produced more promising results, though no clear path to improved GNN performance was found. For this task, maybe less is more.

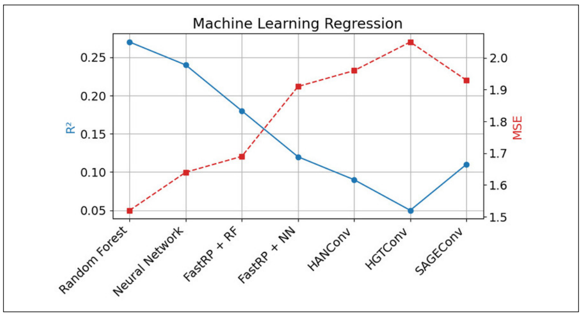
Screenshot of the Neo4j browser showing the knowledge graph
Own presentment



Overview of the data collection, preparation and knowledge graph construction pipeline
Own presentment



Regression results on the machine learning dataset
Own presentment



Advisor
Dr. Shao Jü Woo

Co-Examiner
Mag. Dipl.-Ing. Dr. Kathrin Plankensteiner, FHV - Vorarlberg University of Applied Sciences, Dornbirn, AT

Subject Area
Computer Science, Data Science