Journal Classification Using Graph-Based Machine Learning on a Bibliographic Dataset

Syed Fahad Nadeem (sn07558)

Group 9

Graph Data Science, Spring 2025

Abstract—This project applies graph-based machine learning to classify journals into subject categories using a bibliographic dataset. Utilizing Neo4j Graph Data Science (GDS 2.12.0), we constructed a graph with 126 journals and 30,441 papers, incorporating relationships like PUBLISHED_IN and CITES. A node classification pipeline with FastRP embeddings and logistic regression achieved a test accuracy of 31.6%, though uniform predictions highlighted challenges from a small dataset and class imbalance. The methodology, results, and potential improvements are discussed to provide insights into graph-based bibliographic analysis.

Index Terms—Graph Data Science, Node Classification, Bibliographic Network, Neo4j, Journal Categorization

I. Introduction

Bibliographic datasets offer valuable relational insights, making graph-based machine learning an effective approach for tasks like journal classification. This project focuses on predicting journal categories (e.g., Social Sciences, Natural Sciences) based on citation patterns within a dataset derived from migration research [1]. With 126 journals and 30,441 papers, we modeled the data in Neo4j and applied Graph Data Science (GDS 2.12.0) to train a classification model.

The objective was to leverage network structure and node properties to improve category predictions. However, challenges such as a small dataset, class imbalance, and multi-dimensional journals led to uniform predictions (category 0 for all). This report outlines the methodology, presents results, and discusses limitations and future directions.

II. Data Preprocessing and Graph Modeling

A. Dataset Description

The dataset, sourced from [1], includes:

- Journal nodes: 126 journals with names and publishers.
- Paper nodes: 30,441 papers with Paper_field (e.g., "Sociology;Computer Science").
- Relationships: PUBLISHED_IN (from papers to journals) and CITES (between papers).

Total nodes: 30,567; relationships: 111,283.

B. Data Preprocessing

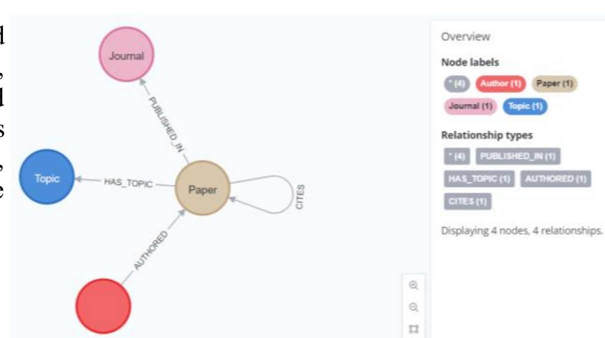
Data was loaded into Neo4j, focusing on Journal and Paper nodes. We cleaned Paper_field by splitting and trimming values (e.g., "Sociology;Computer Science" → ["Sociology", "Computer Science"]). A Cypher query assigned categoryId to journals based on paper fields, mapping to categories like Social Sciences (0) and Engineering & Technology (3). The class distribution showed imbalance: category 0 (37 journals), category 2 (33), others smaller (e.g., category 5: 7).

C. Graph Modeling

The graph model included:

- Nodes: - Journal: Properties: categoryId, category (e.g., "Social Sciences"). - Paper: Properties: Paper_field, categoryId (initially -1).
- Relationships: - PUBLISHED_IN: Paper to Journal. - CITES: Paper to Paper.

The graph was projected into GDS as journal_citation_graph.



III. Methodology

A. Node Classification

The task was to predict categoryId (0 to 6) for Journal nodes based on citation patterns.

1) Graph Projection: We projected journal_citation_graph in GDS with:

IV. Results

A. Model Performance

The model achieved:

- Test Accuracy: 31.6%.
- F1_MACRO: 6.9%.

All predictions were category 0, aligning with the majority class (37/126 journals, 29.4%).

B. Graph Statistics

- Total nodes: 30,567 (126 Journal, 30,441 Paper).
- Total relationships: 111,283.
- Class distribution: 0 (37), 2 (33), 1 (15), 4 (13), 3 (12), 6 (9), 5 (7).

```
Graph Projection Result:
              graphName  nodeCount  relationshipCount
0  journal_citation_graph      30567             111283
```

```
Creating node classification pipeline 'journal_classification_pipeline'...
Adding FastRP feature step to pipeline...
Configuring split...
Adding logistic regression model parameters...
Training the model 'journal_category_model'...
Model Training Result:
```


```
        modelInfo
0  {'classes': [0, 1, 2, 3, 4, 5, 6], 'modelName': 'journal_category_model', 'featureProperties': [], 'modelType': 'N
odeClassification', 'metrics': {'F1_MACRO': {'test': 0.06857142799542856, 'validation': {'min': 0.06211180072913854,
80074264112, 'max': 0.06279434800489936, 'avg': 0.06256683225081328}}, 'ACCURACY': {'test': 0.31578948, 'validation':
 {'min': 0.27777778, 'max': 0.28571429, 'avg': 0.28042328333333333}, 'outerTrain': 0.28971963, 'train': {'min': 0.277
77778, 'max': 0.28169015, 'avg': 0.2803860266666666}}}, 'pipeline': {'featureProperties': [], 'nodePropertySteps': [{
'name': 'gds.fastRP.mutate', 'config': {'randomSeed': 42, 'contextRelationshipTypes': [], 'mutateProperty': 'fastRP_e
mbedding', 'iterationWeights': [0.7, 0.2, 0.1], 'embeddingDimension': 1024, 'contextNodeLabels': []}}]}, 'bestParamet
ers': {'minEpochs': 1, 'maxEpochs': 100, 'focusWeight': 0.0, 'patience': 1, 'tolerance': 0.001, 'learningRate': 0.001
, 'batchSize': 100, 'penalty': 0.01, 'methodName': 'LogisticRegression', 'classWeights': [1.0, 1.5, 1.2, 1.8, 1.6, 2.
5, 2.0]}, 'nodePropertySteps': [{'name': 'gds.fastRP.mutate', 'config': {'randomSeed': 42, 'contextRelationshipTypes'
: [], 'mutateProperty': 'fastRP_embedding', 'iterationWeights': [0.7, 0.2, 0.1], 'embeddingDimension': 1024, 'context
NodeLabels': []}}]}
```


V. Discussion

A. Challenges

Uniform predictions stemmed from:

- Small Dataset: 126 journals limited training data.
- Class Imbalance: Category 0 (37 journals) dominated.
- Multi-Dimensional Journals: Journals with diverse paper fields (e.g., Sociology and Computer Science) complicated categorization.
- GDS 2.12.0 Limitation: Restricted to logistic regression, less suited for imbalanced data.

B. Alternative Approaches

- More Data: Increasing journal count could enhance learning.
- Advanced Models: GNNs (available in newer GDS versions) could capture complex patterns.
- Additional Features: Adding paper citation counts could enrich embeddings.

C. Possible Extensions

Future work could involve clustering papers to infer categories or integrating paper metadata (e.g., titles) for hybrid classification.

VI. Conclusion

This project applied graph-based machine learning to classify journals, achieving 31.6% accuracy but facing uniform predictions due to dataset constraints. The methodology and analysis provide a foundation for future enhancements, such as larger datasets or advanced models, advancing bibliographic research with graph techniques.

References

[1] L. Rothenberger, M. Q. Pasta, and D. Mayerhoffer, “Mapping and impact assessment of phenomenon-oriented research fields: The example of migration research,” Quantitative Science Studies, vol. 2, no. 4, pp. 1466–1485, Dec. 2021. [Online]. Available: https://doi.org/10.1162/qss_a_00163