**Scientific Report on Citation Prediction Project**

**Abstract**

This report presents a comprehensive study on citation prediction, focusing on the task of predicting whether one academic paper will cite another based on various features. The project employs both traditional machine learning methods and advanced graph-based approaches, like Graph Neural Networks (GNNs), to enhance prediction accuracy.

**1. Introduction**

**1.1 Background**

Citation networks are pivotal in academic research, serving as a means to track the influence and interconnectivity of scholarly work. They provide insights into how knowledge is disseminated and how research evolves over time. Citation prediction, the task of forecasting whether one paper will cite another, has significant implications for academic publishing, helping researchers identify relevant literature and understand citation dynamics.

* 1. **Task Definition**

The primary objective of this project is to predict the likelihood of citation between pairs of academic papers based on various features derived from their metadata and citation relationships. This involves analyzing the characteristics of the papers, such as their titles, authors, and publication venues.

* 1. **Significance of the Study**

Understanding citation behavior is crucial for researchers, publishers, and academic institutions. Citation prediction can facilitate literature reviews, enhance recommendation systems, and improve the visibility of scholarly work.

It is in our opinion also very interesting to discover which are the most important factor that make one paper cite another. Discovering also hidden patterns

**2. Literature Review**

Existing methods for citation prediction primarily rely on traditional machine learning techniques, which often face limitations such as feature sparsity and the inability to capture complex relationships within citation networks. Recent advancements in graph-based approaches, particularly those utilizing graph neural networks, have shown promise in addressing these challenges. These methods leverage the structural information inherent in citation networks, allowing for more nuanced predictions.

**3. Methodology**

* **3.1 Data Collection**

The dataset utilized in this project comprises a comprehensive collection of academic papers formatted in a txt file. Each paper includes essential metadata, such as titles, authors, publication years, venues, and citation information. In total, the dataset encompasses 629,814 papers. Initially, it contained six key features: title, authors, year, venue, index, and citations. Notably, a significant portion of the dataset had missing values, with 504,442 entries for citations marked as NaN, 348,734 entries for abstracts also recorded as NaN, 25,916 authors marked as empty and also 98,594 venues marked as empty.

* **3.2 Data Preprocessing**
  + 3.2.1 Data extraction

The data preprocessing phase was essential for converting the raw text dataset of academic papers into a structured format suitable for analysis. The process began with reading the text file containing the papers, which were formatted with specific prefixes to denote different types of information.

The parsing function was designed to read the entire content of the file and split it into individual blocks, with each block representing a separate paper. This was accomplished by using double newlines as delimiters. Each block was then processed to extract relevant metadata, including the title, authors, publication year, venue, index, citations, and abstract. The extraction relied on specific line prefixes: for example, lines starting with “#\*” indicated titles, while lines starting with “#@” denoted authors.

Another dataset was used to add one feature denominated as “categories” which waas then encoded using one-hot-encoding to extract useful information.

* + 3.2.2 Handling missing values

The missing values regarding abstract and authors could be possibly imputed from a cross search online for each missing row, but this was considered to be too time consuming for the time we had.

The missing values for citations were considered correctly empty, as well as the ones for venues.

In the year feature we found 9 papers with a -1 year imputed, we decided to impute the year for the relevant papers and drop the others based on the number of citation they had.

* **3.3 Feature Engineering**

Several features were engineered to enhance the predictive power of the models. These include:

* Title Similarity: A measure of the similarity between the titles of two papers.
* Author Features: Metrics such as the number of common authors and the total number of authors for each paper.
* Venue Features: Information regarding the publication venues, including the number of papers published in each venue.

In addition to these traditional features, we also incorporated features related to the citation graph for the graph-based model. These graph-related features aimed to capture the structural relationships within the citation network, such as node connectivity and the degree of each paper within the graph.

The choice of these features was justified based on their potential impact on citation behavior, as previous studies have indicated their relevance (https://faculty.ucmerced.edu/hbhat/BHRDE2015.pdf). By combining traditional features with graph-based features, we aimed to improve the model's ability to predict citation relationships more accurately.

* **3.4 Data Splitting**

Initially, we performed a random stratified split of the dataset, but we encountered issues with data leakage. This led us to revise our approach to ensure that the training and testing datasets were distinct, preserving the integrity of the evaluation process.

We explored several strategies to address the data leakage issue and ensure a more reliable split:

* Predefined Training and Testing Sets: One approach was to manually assign specific papers to the training set and others to the testing set right from the start. This method aimed to ensure that no paper appeared in both sets, preventing data leakage. By controlling the selection from the beginning, we could more accurately evaluate the model’s performance.
* Paper Pair Creation with Deduplication: Another method involved creating pairs of papers and then ensuring that no paper from the test set also appeared in the training set. By creating these pairs first and then removing duplicates, we minimized the risk of data leakage that could occur if papers from the training set were inadvertently cited in the test set.
* Splitting by Publication Year: We also considered splitting the data based on the year each paper was published. The idea was to ensure that pairs of papers in the training and testing sets had a clear temporal separation. For instance, we could create paper pairs with at least a one-year gap between their publication dates. This would help simulate a more realistic scenario where newer papers cite older ones, but not vice versa.

However, this approach came with its own set of challenges. When splitting by year, we had to be particularly cautious about citations between papers from different time periods. If the citation relationships between training and test papers weren’t properly handled, such as failing to remove citations from training papers to test papers, we risked introducing data leakage. This would allow the model to learn from citation patterns it shouldn’t have access to, ultimately skewing the evaluation.

In the final data preparation for the model, we concatenated the features of two papers, referred to as *paper\_a* and *paper\_b*. We also added a label: 1 if *paper\_a* cites *paper\_b*, and 0 otherwise. Additionally, we incorporated other features that reflect the relationship between the two papers, such as title similarity, which helps capture the semantic connection between them.

Through these approaches, we aimed to create a more robust and reliable dataset that could accurately reflect the citation dynamics, while preventing the model from inadvertently using information from the test set during training.

* **3.5 Rationale for Selecting Only a Subset of the Data**

The initial dataset included a large number of articles, many of which had no citations. This presented a challenge: including articles without citations in the model would have reduced its ability to accurately predict citation relationships. Articles without citations provide no relational information, which is essential for the prediction task.

As a result, we decided to filter the articles, selecting only those with a significant number of citations. This decision was based on the assumption that the most cited articles play a crucial role in the citation network and are more representative of citation dynamics. By including these articles, we were able to:

* Improve the quality of the dataset: By focusing on articles with citations, we created a richer and more informative dataset that includes relational information.
* Build a more meaningful model: Concentrating on articles with citations enhanced the model's ability to capture citation relationships and reduced the "noise" caused by articles without connections.
* Ensure a fair comparison: Our goal was to develop a model that could be evaluated on a consistent basis, using articles that are sufficiently relevant within the citation network.

Although this approach introduces a filter, it did not compromise the generalizability of the model. Instead, it allowed us to focus on a representative and informative subset of the dataset. In real-world scenarios, a similar strategy could be used to highlight key articles that form the core of an academic network.

* **3.6 Model Selection**

For the classification task, various models were considered, including traditional machine learning techniques and more advanced graph-based methods. The primary models evaluated were:

* + 1. **Random Forest**: The Random Forest algorithm was tested on different subsets of the dataset using a range of feature combinations. This approach helped assess how well traditional features, such as title similarity, author overlaps, and venue information, could predict citation relationships. In addition, a modified version of the Random Forest model was created by incorporating graph-based metrics to enhance prediction accuracy. These graph metrics were derived from the citation network, which was constructed to capture the structural relationships between papers. The specific graph metrics used included:
  + **In-degree**: The number of incoming citations a paper receives.
  + **Out-degree**: The number of papers a given paper cites.
  + **Betweenness centrality**: A measure of how often a paper lies on the shortest path between two other papers.
  + **PageRank**: A measure of the importance of a paper based on its citation relationships, akin to the algorithm used by Google Search.
    1. **Graph Neural Networks (GNNs)**: To leverage the inherent structure of the citation network, Graph Convolutional Networks (GCNs) were also explored. These models were tested with different combinations of data features, including both traditional metadata and graph-based features. GCNs have the advantage of learning directly from the graph structure and are particularly suited for tasks like citation prediction, where the relationships between papers are crucial. The experiments with GNNs helped us assess whether they could outperform traditional machine learning models by better capturing the complex interactions between academic papers.
    2. **Justification for Model Choice:**  The selection of the final model was based on preliminary results from both Random Forest and GNN evaluations. While Random Forest performed adequately with traditional features, its ability to incorporate graph metrics significantly improved its accuracy. However, the GNN models demonstrated superior performance in capturing the complex citation dynamics within the network, as they were designed to work directly with graph data.

Ultimately, based on the performance metrics, including precision, recall, and F1-score, the GNNs were chosen as the final model for this study. The ability of GNNs to better understand and predict citation relationships, given their architecture and the richness of graph-based features, was a decisive factor in this choice.

**4. Experimental Trials**

* **4.1 Initial Trials**
  + Overview of the first attempts using traditional machine learning models.
  + Discussion of the results and challenges faced (e.g., imbalanced dataset, data leakage).
* **4.2 Graph-Based Approach**
  + Description of the transition to a graph-based approach.
  + Explanation of how the citation graph was constructed and the features derived from it.
* **4.3 GNN Implementation**
  + Overview of the Graph Neural Network architecture used.
  + Discussion of the training process and evaluation metrics.

**5. Results**

* **5.1 Model Performance**
  + Presentation of the results from the Random Forest model and the GNN.
  + Comparison of performance metrics (e.g., precision, recall, F1-score, AUC-ROC).
* **5.2 Feature Importance**
  + Analysis of feature importance for the Random Forest model.
  + Discussion of how graph-based features improved model performance.
* **5.3 Visualization**
  + Include visualizations of the citation graph, degree distributions, and model performance metrics (e.g., confusion matrix, ROC curve).

**6. Discussion**

* **6.1 Interpretation of Results**
  + Discuss the implications of the results and what they reveal about citation behavior.
* **6.2 Limitations**
  + Acknowledge any limitations of the study (e.g., dataset size, feature selection).
* **6.3 Future Work**
  + Suggestions for future research directions, including potential improvements to the model and additional features to explore.

**7. Conclusion**

* Summarize the key findings of the project.
* Reiterate the importance of the study and its contributions to the field of citation prediction.

**8. References**

* List of all academic papers, articles, and resources cited throughout the report.

**9. Appendices**

* Additional material, such as code snippets, detailed results, or supplementary figures.