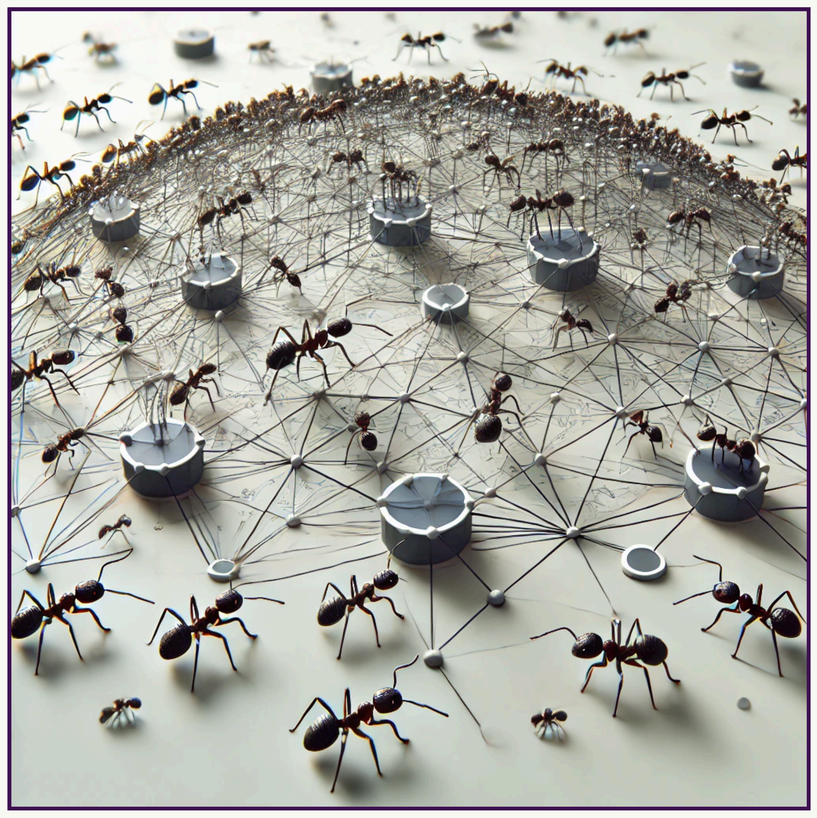
Citation Prediction using Ant Colony based Multi-Level Network Embedding

AUTHORS
Itamar Kraus
Yoni Azeraf

SUPERVISOR
Dvora Toledano-Kitai



Abstract

This project addresses the challenge of academic papers including citations that are either weakly relevant or lack substantial significance. Such citations negatively impact the quality and relevance of research , as well as the ability of readers to navigate the vast landscape of academic information. Existing approaches, such as manual screening and expert reviews suffer from subjectivity and limited accuracy. To tackle this issue, this project introduces a novel approach based on Ant Colony based Multi-level Network Embedding (ACE), drawing insights from "ACE: Ant Colony based Multi-level Network Embedding for Hierarchical Graph Representation Learning" [11]. This project introduces a different approach. Specifically, it constructs a citation graph and applies an Ant Colony Optimization (ACO) -based algorithm [7] to identify and assess the relevance of citations. This method aims to enhance both the accuracy and objectivity of citations relevance evaluation in academic research.

Proposed Model

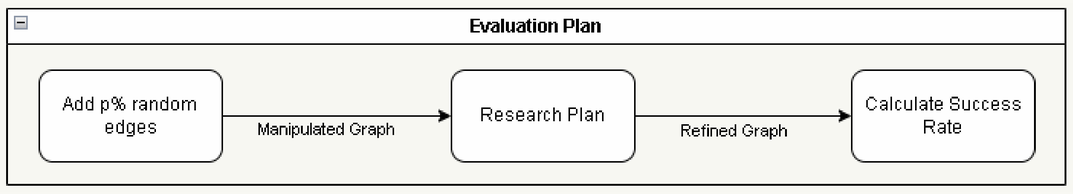
Detects irrelevant citations using a ACE-based approach on the CORA dataset.

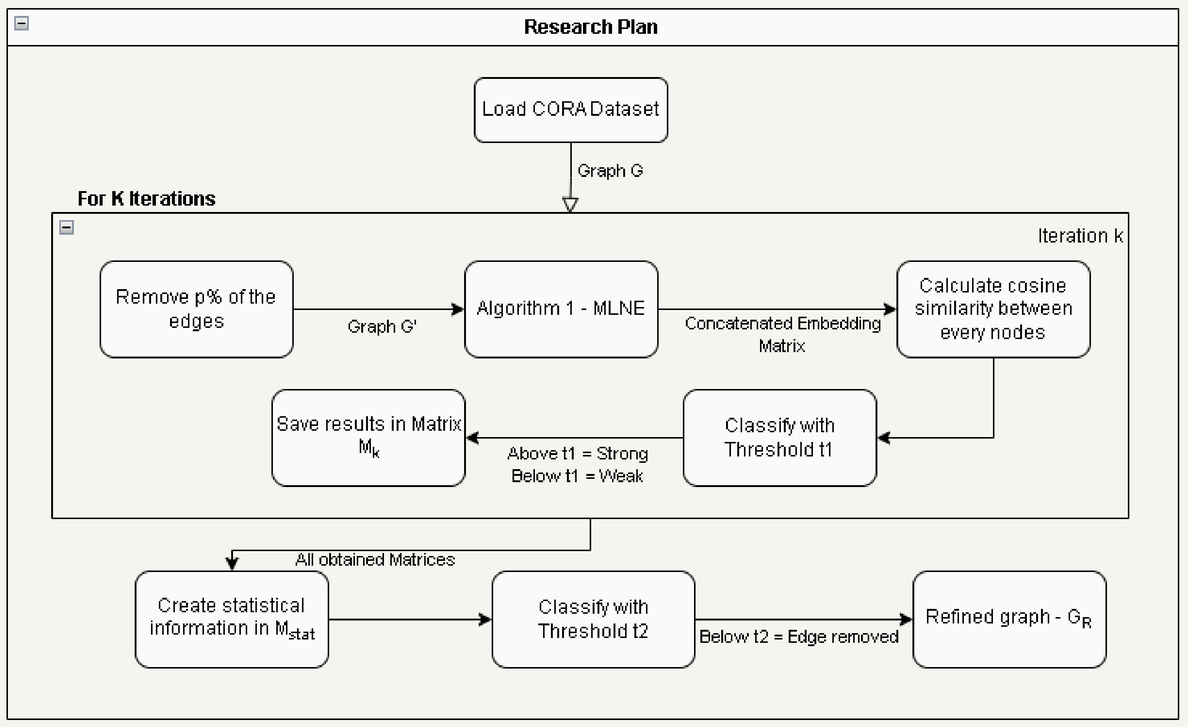
- Runs for K iterations, removing a predetermined percentage of edges in each iteration.
- Processes the modified graph using the MLNE algorithm and node2vec for multi-layered embeddings.
- Computes cosine similarity between node embeddings to determine closeness.
- Stores closeness results in a binary matrix and constructs a statistical matrix over multiple iterations.
- Applies a predefined threshold to classify relationships as strong or weak.
- Removes weak connections (irrelevant citations), refining the graph to retain only significant, reliable edges.

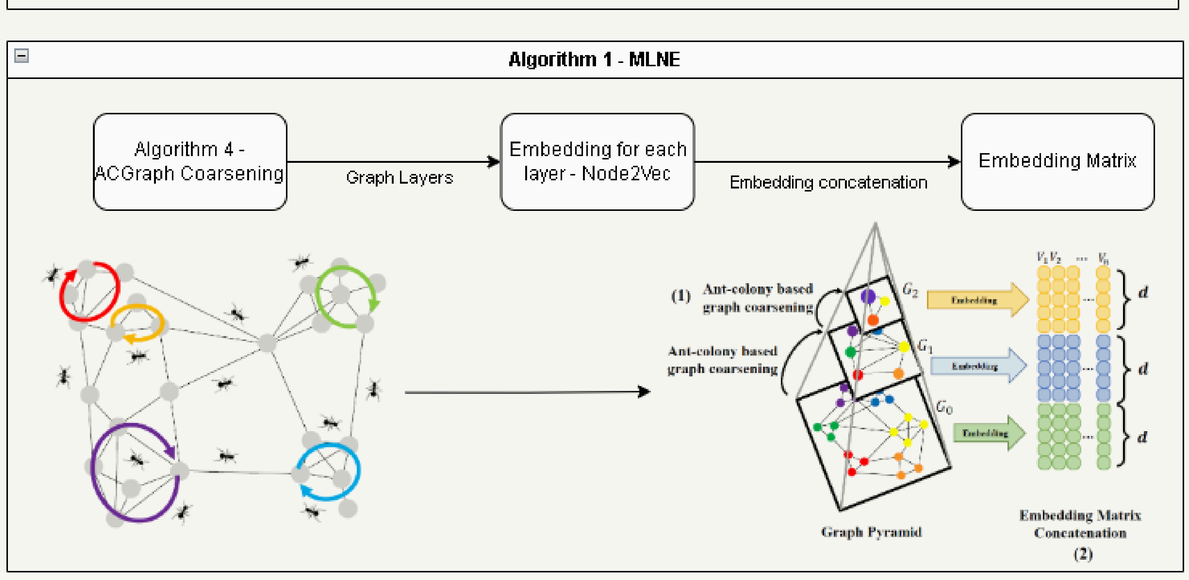
Model Evaluation

Randomly add edges to the CORA graph, check if detected using the Proposed Model.

- A predetermined percentage of edges was randomly added to the CORA-based graph, saved as *E'*.
- The proposed method was applied to the modified graph, and the output was analyzed.
- Since the added edges were random, they likely connected nodes with weak relationships.
- The algorithm was expected to identify and remove these weak edges.
- Success was measured by calculating the percentage of edges from *E'* that were absent in the final output.
- The evaluation was performed multiple times, and results were averaged to ensure accuracy.
- This approach assessed the model's overall effectiveness in detecting and removing irrelevant edges.







RESULTS

Model Evaluation Results

To validate the proposed method, we tested various hyper-parameters with the following assumptions: *t1 = 0.9*, edge removal *30%*, *30* iterations, *8* pyramid layers, and *alpha = 0.5*.

Embedding vector size	T1	T2	Edge removal (%)	Iterations(K)	Alpha	Pyramid scales	Success Rate(%)
128	0.9	50	30	30	0.5	8	84.396
256		40					98.989
512		35					98.926
1024		30					96.968

Over 95% of artificially added edges were successfully removed, aligning with expectations.

Model Results

After validating the model with good results, we tested the original CORA graph without artificial edges. Based on optimal results from prior experiments, we selected an embedding vector size of 256 and a *Threshold t2* of 40% for these experiments.

Embedding vector size	T1	T2	Edge removal (%)	Iterations(K)	Alpha	Pyramid scales	Weak edges classified(%)
256	0.9	40	30	30	0.5	8	51.25
			40				53.22
			50				62.59

The experiments varied in the percentage of edges randomly removed in each iteration: 30%, 40%, and 50%. A clear relationship emerged: the higher the percentage of edges removed, the higher the percentage of edges classified as weak. This suggests that processing a smaller graph at each step allows for more effective identification of weak edges.

Conclusion

Previous researches indicates that about 66% of citations are irrelevant, which aligns with our experimental results of ~63%. While future experiments may refine this further, our current results are the best achievable within the given time constraints.
The distribution analysis shows that most edges were not recovered, with the highest percentage at '0', reinforcing that a significant portion of the CORA graph consists of weak connections.