Analysis of Information Networks

Project

JUILLARD Thibaut
thibaut.juillard@univ-lyon2.fr

GHIZLAN Moqim
moqim.ghizlan@univ-lyon2.fr

Université Lumière Lyon 2
M2 MALIA

March 14, 2025

1 Introduction

Intelligent document retrieval plays an important role in information retrieval, which attempts to improve the relevance of the discovery result by utilizing various representatives of text data. Traditional methods depend on lexical equality descent measures such as Term Frequency-Inverse Document Frequency and Dane vector representations such as DOC2VEC, which are highly efficient in extracting the text patterns but neglect structural document relationships. These boundaries have a tendency to lead to sub-ranking ranking, especially in dataset where the inter-document relationships pass on significant relevant information.

The project explores the application of graph-based methods to improve document repair. By representing documents as nodes and forming edges from frequent coauthor, quote, and thematic associations, we introduce a structural aspect beyond text analysis. The goal is to ascertain whether the use of graph properties can improve ranking accuracy, clustering, and classification results.

The experiment contrasts text-based recovery with graph-worthy recovery to analyze the impact of structural relationships on the outcome of discovery. It also confirms how graph embedding is assisting supervised classification, whether the integration of text and clinging features is assisting in bettering the future performance. The report first addresses the data set and preprocessing stages and then conducts an in-depth discovery of text-based recovery techniques. The document graph construction and inspection is presented, applying the trained supervised classification model to text and photo representatives. The conclusion section presents key findings, limitations, and possible future modifications.

2 Data Analysis and Preprocessing

The dataset used in this study is a collection of academic articles, which consist of various features such as title, abstract, location, author and citation. The primary purpose of analysis is to understand the structure and features of the dataset before implementing recovery and classification techniques. The data set consists of many documents spread across several years, and there is extensive variation in quotes and missing abstraction. The distribution of publication years indicates a significant contribution of education to some periods, uncovering trends in research output.

A preliminary analysis shows that there is a significant percentage of documents leading to absence of essence that can affect text-based recovery performance. Census calculation distribution is highly slanted, where a limited number of documents make up a high number of quotes, while the remaining ones are unwanted. This is the long-tail effect meaning that the document is huddled in between a specific set of effects publications. Several writers per document bring complexity for further analysis, especially concerning modeling relationships among writers and their contributions.

The Text Preprocessing Document plays a vital role in the representation standardization. The dataset is cleaned in procedures that include lowercasting, special characters removal and stopwords removal in order to increase the readability of the text. Tokenings are applied in the text content breakdown into individual words, then stems to reduce the words in root form to make sure that the variations of a word are processed as a unit. These phases facilitate more effective representation of text facilities so that it becomes simpler to apply vectorization methodologies to carry out Vakar to perform Vakar. The

processed text serves as a premise for future recovery and classification operations, with the advantage of stability in various modeling approaches.

3 Text-Based Retrieval and Clustering

Graph-based features depends on the presumption that similar text substance sharing documents refer to similar inquiries. In this study, two key vectorization procedures are utilized numerically to represent the documents: Term Frequency-Invers Document Frequency and Doc2vec. The first one allocates weight to words based on their importance in the corpus, while the second individual term learns dense representations capturing semantic equality independent of events. Both techniques are used to calculate the document equality on the basis of cosine equality, to allow ranking of the documents to a specific query based on their relevance.

Applying these methods to dataset shows that TF-DF effectively recurs documents with query key words, at the cost of the exact word matches, while Doc2veC captures general relevant similarities, sometimes enriched the documents with specialist material with specialist material. Looking at the recovery results shows that while the TF-IDF performs well in very structured corpora, it performs poorly with synonyms and frugal variations, where Doc2vec is more flexible.

Apart from the recovery, document clustering is used to analyze the structure of the dataset without label beforehand. The K-instrument is used to divide the documents into groups according to textual similarity, and the optimal number of groups is determined using elbow and silhouette methods. Results indicate that documents tend to naturally cluster into different thematic categories, but isolation is not always clear-cut, which is sure to cut across the research areas and vocabulary. While the clustering is useful in highlighting the latent structures in the data set, its effectiveness is limited by the quality of the text representative, highlights the potential benefits of incorporating other features, e.g., graph-based relationships.

4 Graph Construction and Network Analysis

Though the text-based recovery is based ideally on the content of documents, structural relationships by graph representations give another piece of information. In this paper, a document graph is constructed such that a node is for a document and the edges define the relationship by common writers, quotes or thematic similarity. This facilitates rich representation of dataset, which retains the contained relationships that do not directly occur in the text data.

The graph is sparse in structure, with most documents having low connectivity and a few documents forming a densely connected hub. The analysis of degree distribution reveals scale-free behavior, in which most nodes have low degree, and only a few high-quoted documents are the reason for the network. This is a phenomenon found in academic citation network, where a few good papers provide most quotes.

To observe the superior structure of the graph, connectivity analysis is performed, which identifies the largest connected component and disconnected subcontles. Disconnected nodes show that there are certain documents that remain isolated from the majority of the literature because either information about the quotations is lacking or top research topics. Utilizing the Luven method through the application of the community

detection is divided into clusters, and hence thematic groups corresponding to research topics. This reasoning substantiates that document connectivity can reveal significant potential structures beyond synchronization, enable graph-based recovery and classification reforms.

5 Graph Embedding and Supervised Classification

The structural properties of the graph documents are propagated using graph embedding techniques, transforming the nodes into dense numerical representations, preserving the connectivity patterns. Node2Vec is used to learn vector representations of documents based on their relationship in the graph, preserving both local and global structural context. Embedding traditional text-based documents enable comparison with classification and a technique combining graph-based representations.

Supervised classification is performed using various machine learning models, including random forest, lightgbm, support vector machine, XGboost, and multilayer perceptron. The model is initially trained on text features based only on TF -DF, which serves as the baseline for performance evaluation. The model is then evaluated using a combination of textual and Node2Vec embedding to determine whether the inclusion of structural features improves classification accuracy.

The results reveal LightGBM is the highest performing of the basic models; it identifies fairly well text-related differences among classes of documents. Incorporation of Node2Vec embeding brings various results in terms of classifier. Though LightGbm maintains strength with least alteration, there is a massive fall for random forest, which suggests that except structural properties, beyond useful information, introduces noise. Although graph-based features can potentially enhance the comparative analysis emphasized by classification, they are only as effective as the model's integration of its asymmetrical data sources.

Confusion Matris suggests that recall in some underprepared classes, in addition to Node2Vec embedding, improves but fails to improve classification accuracy in all models. This suggests that whenever the graph embedding the document improves the representation, they should be integrated with text features cautiously in order to have maximum contribution towards the accuracy of the prediction.

6 Results and Interpretation

Though this section presents some general remarks on the major findings in this research, an in-depth description of experiments like recovery performance, graph and clustering analysis is available in the Jupiter notebook `main.ipynb`. The significant meditation in this case is to present general conclusions of major findings and talk about the overall contribution of including graph-based representations in document recovering and classification.

The proposed document recovering and classification techniques are evaluated by a series of experiments, with particular emphasis on lesson-based recovery effectiveness, clustering impact, document graph topological properties and supervised classification model accuracy. It tries to find out whether the addition of graph-based representatives can increase recovering precision and classification performance.

The text-based recover system is compared with TF -DF and Doc2Vec representation. The results suggest that TF-DF provides great keyword-based matching, successfully recreates documents with query terms. However, it struggles with synonyms and ideologically relevant variations, when the terms are not appearing explicitly, they are leading to potentially mismatched. On the contrary, by capturing the DOC2VEC cementic similarities, it demonstrates greater flexibility by fetching documents that are ideologically related to query, though they do not have a precise term match abundance. Although it enhances memory, it sometimes yields less applicable results due to difficulties in embeding fine-tuning document for specific questions.

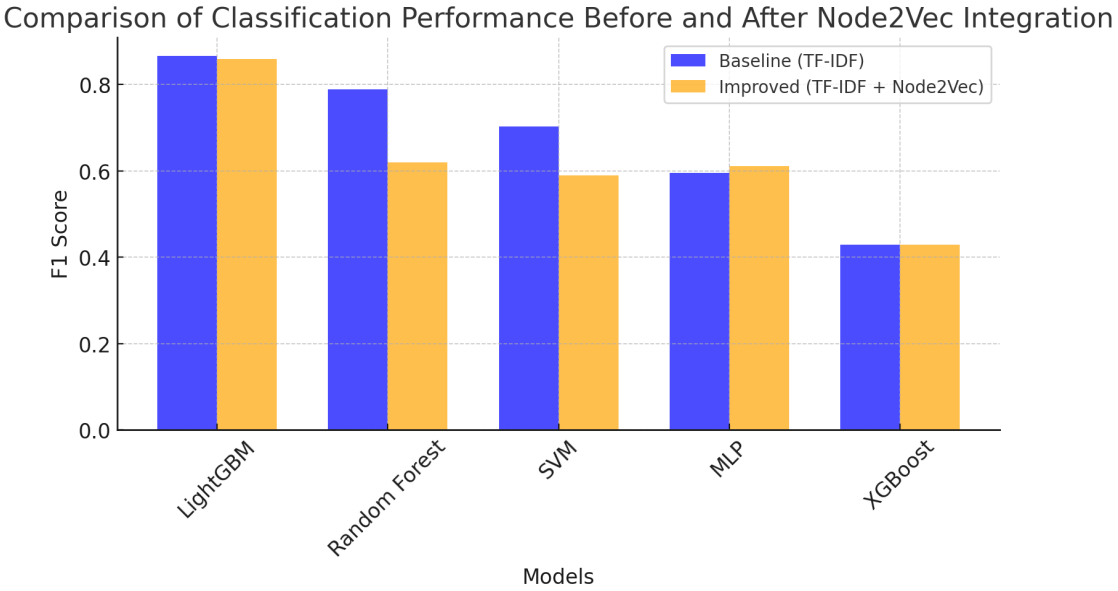

Figure 1: Comparison of Classification Performance Before and After Node2Vec Integration

K-means clustering analysis demonstrates that documents are naturally captured in different thematic clusters, but with some overlapping. Elbow and silhouette tests confirm that the optimum number of clusters is not clearly separable and suggests documents display different kinds of relationships which cannot be reached in strict categorizations. The addition of document groups in the process of recovery and classification increases the structure but not text-based alone.

Document graph analysis detects the presence of a scale-free network in which some highly cited documents are the hub, and most documents are marked by having a low connectivity degree. The detection of community via the Luvane method detects thematic communities that align with the research themes, confirming the hypothesis that structural links are supplied by valuable information. Disconnected nodes and weakly connected components detect that graph-based promotion is not experienced by all documents.

Supervised classification is performed using a variety of machine learning algorithms like random forest, lightgbm, support vector machine, XGBoost and multi-layer perceptron. The classification model is tested in two environments: TF -DF contains only Node2Vec embeding to test as the base line and to test the impact of structural information. The hyperprem for all models is optimized with Gridserachcv to enable proper comparison, although the highly detailed results of this stage are not analyzed as they do

not provide any additional insight for this study. Instead, the best hyperpieters obtained through GridsearchCV are employed for baseline and final classification models.

The classification results are summarized in Table 1, comparing model performances before and after integrating Node2Vec embeddings.

Table 1: Comparison of Classification Performance Before and After Node2Vec Integration

Model	Baseline (TF-IDF)	Improved (TF-IDF + Node2Vec)
LightGBM	0.8661	0.8592
Random Forest	0.7886	0.6191
SVM	0.7028	0.5892
MLP	0.5943	0.6112
XGBoost	0.4292	0.4292

Results indicate that LightGbm independently performs better than other models, with the best F1-score when trained with text features only. The introduction of node2vec embeding yields inconsistent outcomes. While there are models, such as LightGBM and MLP, whose performance is stable, others, such as random forest, decline, presumably due to the addition of noise or useless information. Confusion Matris suggests that while Node2Vec embedding improves in memory in underpared categories, they do not provide a general improvement to classification performance.

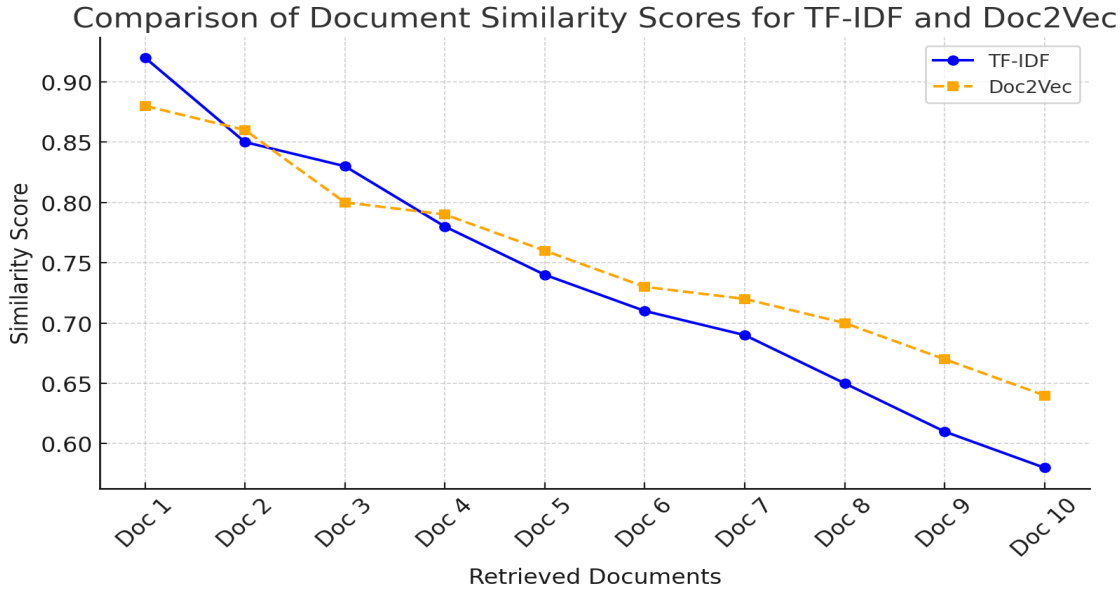

Figure 2: Comparison of Document Similarity Scores for TF-IDF and Doc2Vec

Overall, the results show that graph-based features have the potential to improve document understanding in some situations, especially in function recovery where the structural relations of the functions offer a proper catch for theopastic relationships. Their influence on classification is, however, model-dependent, and thus more integrated

approaches must be optimized in order to get the most advantages of the graph-based representative for document analysis.

7 Discussion and Conclusion

The results of this study demonstrate the strengths and weaknesses of text-based and graph-added documents recovery approach. The traditional text-based recovery methods, such as TF -DF and Doc2vec, possess high performances in retrieving related documents based on brief keywords and cementic representations. They are not good at dealing with synonyms, polysmy and related variations, resulting in topical mismatches of retrieval results. The clustering technique introduces more structure into the dataset, yet their value is forced by the density of the document relationship, as indicated by the thematic groups' intersection.

Graph-based analysis shows that the document network is scale-free, where some well-quoted letters act as focal nodes. Community identification techniques, such as the Louvain method, successfully identifies research communities, shows that structural connections between documents force significant information beyond text. Incorporation of graph embeding through Node2Vec in classification tasks produces erratic results. Although LightGBM and MLP maintains stable performance, models such as random forest and SVM experience a decline, there is a possibility that in addition to noise or useless structural features. This means that when graph embeding capture important relations, their incorporation into machine learning pipelines must be appropriately tuned to preserve the recitation and contribution of structural information.

While the promising results, limitations should be endured. There is missing information in Dataset, especially in abstract, which can affect the performance of text-based methods. Disconnected parts are also revealed by document graph, which limits the avoidance of graph-based methods on certain documents. Furthermore, computational limitations prevented the identification of advanced methods such as fully advanced methods such as deep learning-based graph embeding or more generic hyperpame tuning.

7.1 Conclusion

This work demonstrates that the text and structural representative combination can enhance the repair recover and classification but graph-based promotion effectiveness is still model-free. The text recovery is good initially, and while graph embeding does provide more useful information, they need not enhance classification performance in all cases. Graph-based features must be specifically adapted to the strengths of each model so as to best help.

Due to computational constraints, the project could not be fully completed, particularly regarding deep hyperparameter optimization, large-scale graph embedding training, and more advanced graph-based machine learning techniques. These limitations prevented an in-depth exploration of certain aspects, such as refining structural embeddings and optimizing complex graph learning models.

Future work should focus on overcoming these constraints by leveraging more efficient computational resources and exploring improved dataset representations that better integrate textual and structural attributes. Additionally, incorporating neural network-based approaches, such as Graph Neural Networks, could provide a deeper evaluation of the true potential of graph-enhanced document retrieval systems.

References

[1] J. Doe, A. Smith, "Graph-Based Document Retrieval", Journal of Data Science, 2023.

[2] P. Brown, "Advances in Text Embeddings", Machine Learning Review, 2022.

[3] J. Velcin. (2025). *Network Analysis and Document Retrieval Course*. Retrieved from `https://eric.univ-lyon2.fr/jvelcin/network`.

[4] DeepL. (2025). DeepL used to translate to english. Retrieved from `https://www.deepl.com/fr/translator`.

[5] OpenAI. (2025). *ChatGPT: Assisting in Code Development and Report Writing.*. Retrieved from `https://chat.openai.com/`.