ICIC Express Letters
Volume **13**, Number **10**, October **2019**

ICIC International ©2019 ISSN 1881-803X
pp. 883–889

EXPLORATORY ANALYSIS OF LEGAL CASE CITATION DATA USING NODE EMBEDDING

Shreyansh Lodha and Rupali Wagh

Department of Computer Science
CHRIST (Deemed to be University)
Bangalore, Karnataka 560029, India
shreyansh.lodha@mca.christuniversity.in; rupali.wagh@christuniversity.in

Received March 2019; accepted June 2019

Abstract. *Legal case citation network is primary tool to understand mutable landscape of the legal domain. These networks are also used to study legal knowledge transfer, similar precedents and inter-relationship among laws of a judiciary. These networks are often very huge and complex due to the multidimensional texture of this domain. In recent years, network embedding using deep learning emerges as a promising breakthrough for analyzing networks. This paper presents a novel approach of learning vector representation for a legal case based on its citation context in the network using node2vec algorithm. These vector embedding are further used in understanding similarities between cases. Paper highlights that the tSNE reduced representation of the obtained vectors facilitates visual exploration and provides insights into the complex citation network. Suitability of node embedding for application of machine learning algorithm is demonstrated by clustering the node vectors for finding similar cases.*

Keywords: Network analysis, Legal citation network, Graph embedding, Node embedding, Node2vec

1. **Introduction.** A court judgment in a common law system is considered to be repository of legal knowledge and its interpretation. A judgment typically contains multiple sections like case history, counsel's argument with citations to similar precedents, judge's comments and the final verdict. Citations to relevant laws and precedents are very crucial for a legal professional as they constitute the basis for argumentation. Citation networks with cases and law represented as nodes and citations among them represented as edges are popularly used for exhibiting legal relationships. Citation links are instrumental in extracting hidden patterns [1,2]. A pioneering work [3] in this direction explains the significance of simple statistical metrics like degree distribution, centrality and betweenness can be instrumental in understanding the complex reference structures of Canadian judiciary. Subsequently there have been various studies highlighting the relevance of network analysis in the domain of law and even unstructured information from judgment document is studied by researchers using networks [4,5]. Finding most relevant precedents using citation links is one of the most researched problems in legal domain [6,7]. Going beyond statistical measures, dispersion in citation network [8] is used for finding landmark judgments and relevant precedents.

Indian Judiciary is hierarchical in nature with supreme court being at the top which is followed by high courts and district courts. Thus, on any single day the number of legal matters presented in various courts of India is in few thousands which also generated thousands of citations to relevant precedents and laws. This makes citation network for Indian court cases extremely huge and complex. Existing approaches for legal citation

DOI: 10.24507/icicel.13.10.883

analysis are still predominantly based on degree statistics and network structural properties. Also, these approaches focus on specific sub-domain of law [8]. Network embedding with the help of deep learning techniques, is used for learning lower dimensional representations for graph, nodes as well as edges. Thus, complex and multidimensional network data is transformed into a set of features by preserving the node-neighborhood context. This transformed data is suitable for application of machine learning algorithms and can further be used for predictive, prescriptive and descriptive learning tasks. Node2vec is a framework which learns lower dimensional representation of a node by preserving the neighborhood properties by simulating biased random walks. In this paper we present application of node2vec algorithm for learning features of legal case citation network for cases heard in Indian High Courts. Every case in the network is modeled as a node and the citations made in the case are the edges which form the neighborhood of the case. The algorithm node2vec generates vector embedding for node by preserving its citation neighborhood and thus every case is represented as a vector facilitating the application of proximity measures and visual exploration of data. Application of clustering, unsupervised machine learning technique is further used, on vector representation of nodes for obtaining more insights into similar cases. The results obtained highlight the suitability of this approach for analyzing legal citation network.

This paper is organized as follows. Section 2 talks about previous work done in the domain of graph analysis and subgraph sampling. The section further highlights the significance of graph embeddings in this domain. Section 3 describes the experimental setup followed during the course of the study. It elaborates the data set construction and steps of the proposed approach. Section 4 contains discussion on the results and interpretations and Section 5 concludes with a note on future work that authors intend to take forward.

2. **Related Work.**

Graph is a complex structure which consists of nodes and edges. It is arguably one of the best structures for describing relationships in many domains like social media, biology, chemistry, Internet topology and collaboration network [14]. Graph structures are used widely. One of the major challenges in using graph structures is the analysis of information/data. Most of the real-world graphs are very huge and complex. The nonlinearity of graph structures makes it difficult to analyze the data [12].

2.1. **Subgraph sampling.**

Subgraph sampling is widely used to reduce the size which in turn reduces the complexity of a graph. It is a method which allows us to reduce the size of the original graph, but the major challenge faced in this approach is to ensure that the subgraph achieved is a “good” subgraph, which is able to preserve the properties of the original graph. Most important graph properties are in-degree, out-degree, hops and clusters [12,13]. Feasibility of a “good” subgraph can be determined by appropriate sampling algorithm and knowledge of the graph. Unbiased algorithms like Metropolis-Hasting Random Walk (MHRW) are likely to perform poor in high-degree nodes in loosely interconnected network than Simple Random Walk (SRW) (Biased) but MHRW is likely to perform better than SRW with high-degree nodes in highly interconnected network [16]. Graph structures can be very large and to extract subgraph from such large graphs requires a sound domain knowledge and deep understanding of all the algorithms. Suitability of a subgraph sampling algorithm is dependent on graph properties and its structure and thus, algorithm selection is a non-trivial problem. Additionally, since data is explained as links and nodes, direct application of generic learning algorithms like prediction, and clustering is cumbersome for graph data which has been one of the major limitations in graph and networks analysis field.

2.2. **Embeddings.** In recent years embedding or vector representation of complex data has gained huge popularity especially for complex data types like graph and text. Embedding in the context of machine learning means for every point xi in high dimension space, we are finding the point xi' in lower dimension space and the same goes for graph embeddings, it is mapping of network to a vector space while trying to preserve the relevant properties. There are various methods to generate graph embeddings like deep learning, graph kernels, and matrix factorization [15]. One of the most commonly known frameworks for generating embeddings is word2vec, which generates embeddings for large textual content also known as corpus. Word2vec can be considered as a framework which exploits word neighborhood over large corpus of text. Using a simple neural network architecture, numeric dense representation of words in the vocabulary are generated. Word2vec contains 2 different models CBOW (Continuous Bag of Word) and skip-gram, with two training methods, negative sampling and hierarchical softmax. **Skip-gram**: works well with small amount of the training data, and represents well even rare words or phrases. **CBOW**: several times faster to train than the skip-gram, slightly better accuracy for the frequent words [10,11]. These methods are proven to be performing better for huge data.

2.3. **Node2vec.** Graph embeddings use similar approach to generate vector representations for various sub-parts of graph, namely nodes, edges, subgraphs and the graph itself [11]. Conceptually node2vec is a framework borrowed from word2vec (skip-gram) which learns lower dimension representation of a node by preserving the neighborhood properties by simulating biased random walks [9], the unique feature of node2vec is sampling strategy which allows to generate corpus from group of directed acyclic graphs, and then we input that corpus to word2vec which generates embeddings for us. These embeddings are extremely useful in the domain of network analysis and predictions, whether it is social media network or chemical structure of a compound [12,13].

3. **Experimental Setup.**

3.1. **Data description.** Case citation dataset used in this paper is obtained from the online legal database [17]. For this study, Information Technology Act 2000 is considered to be the domain of law under consideration and citations of all the cases under the said act conducted in different high courts in different states of India. Thus, in the dataset a case represents a node and the other cases and citations which were referred during the proceedings of the case form links of our citation network. Dataset consists of cases from different high courts in India, it contains total of 1843 unique nodes and 3769 unique edges, with average degree being 4.0901, and each edge holds equal weightage. The graph is made undirected as our aim is to find similar context between cases irrespective of the chronological order in which, case was originated.

3.2. **Proposed approach.** Steps employed during the course of this study are described below.

Step 1: Representation of citation data as citation network

Dataset was mapped to network structure by creating nodes of cases and connecting nodes based on citation. Basic data cleaning was applied to removing cases without any citation. Every node is given unique label.

Step 2: Node2vec to obtain node embedding

The node embeddings were generated by the node2vec algorithm after passing the preprocessed dataset. Parameters: In this study we used following default values for all hyperparameters.

Number of output dimensions = 128, walk length = 80, number of walks = 10, window = 10, minimum count = 1, word batch = 4.

Step 3: Case similarity estimation

Node2vec internally uses word2vec, which gives the output in the form of embeddings and model. This model was used to find similarity between the citing cases.

Step 4: Application of tSNE for 2D visualization of network data in new dimension space obtained through embedding

tSNE, T-distributed Stochastic Neighbor Embedding, is one of the most popular machine learning algorithms used for nonlinear dimensionality reduction. We used this algorithm to reduce the dimensions to 2. Since tSNE works on the principle of preserving proximities among the data objects, the scatter plot of reduced data gives clearer insights.

Step 5: Grouping of similar cases using reduced vector representation of case nodes

The two-dimensional output generated by tSNE algorithm was used as input for, Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The algorithm performs density-based grouping-based grouping of data. Ten distinct clusters were generated by the algorithm.

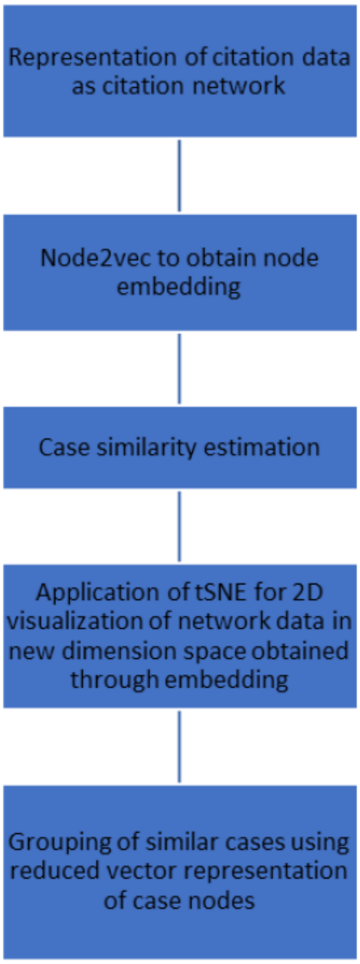
Representation of citation data as citation network
Node2vec to obtain node embedding
Case similarity estimation
Application of tSNE for 2D visualization of network data in new dimension space obtained through embedding
Grouping of similar cases using reduced vector representation of case nodes

FIGURE 1. A flow chart of the complete process

4. Result and Discussion.

4.1. Citation based similarity.

Table 1 shows case similarity values as estimated by the models. Preliminary probe in these values reveals that cases with a greater number of similar legal issues/legal disputers citing the same laws and precedents are given high similarity value by the algorithm. It is observed that the algorithm tends to perform better in terms of similarity estimation for cases with less citations over those with a greater number of citations.

Table 1. Case similarity values returned by node2vec

Case and the legal dispute in the case	Similar case and the legal dispute in the case	Similarity rate (between −1.0 to 1.0)
47 Fraudulent online trading	571 Fake email	0.426769614219666
687 Password or digital identification misuse	1803 Tempering with digital documents, outraging modesty	0.450429111719132
153 Cheating, misleading court, influencing witnesses	1790 Cheating, breach of trust	0.514270901679993
693 Fraudulent transactions	1805 Fraudulent transaction	0.758301973342896
392 Cheating	682 Cheating	0.9224534034729

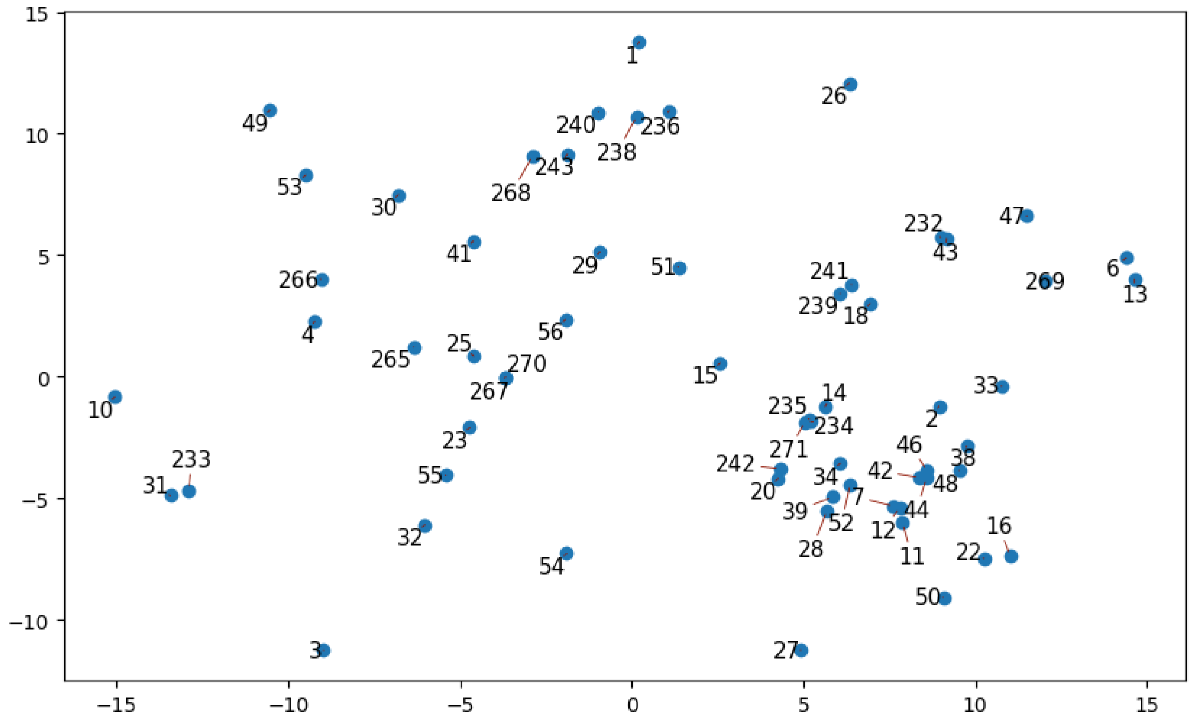

Figure 2. A plot of values generated from tSNE of all the citing cases

4.2. **Data visualization after tSNE dimensionality reduction.** Figure 2 depicts citation network for select nodes (cases) before and after application of node2vec. As it can be seen in the figure, vector representation of nodes and subsequent dimensionality reduction using tSNE algorithm makes the data simpler facilitating visual interpretations.

4.3. **Grouping of similar cases using DBSCAN clustering algorithm.** Figure 3 shows clusters generated by DBSCAN. Preliminary investigations into cluster membership of cases give insights into similar legal concepts handled in the cases. Prominent legal concepts of every cluster identified after human inspection are listed in the table cluster details. Since this is a pilot study, a detailed discussion on accuracy and cluster evaluation

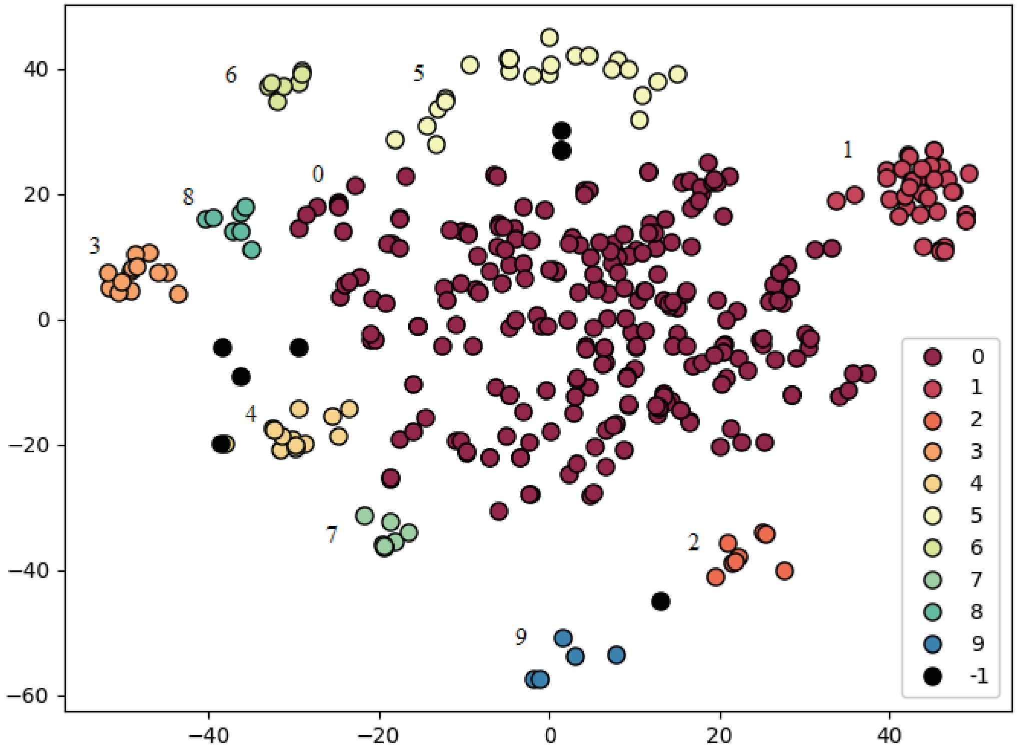

FIGURE 3. (color online) Clusters generated by DBSCAN algorithm

TABLE 2. Case type in each cluster

Cluster number	Prominent legal concept
0	Money laundering, Dishonesty
1	Misuse of digital equipment, Digital evidence, Harassment
2	Posting obscene content on social media, Offensive messages, Cheating using online digital means firms organizations
3	Receipt of electronic records e-signature, Message transmission, Copyright
4	Breach of trust, Offense by clerk or servant dishonesty, Integrity, Religious sentiments, Atrocities against specific community
5	Exchanging obscene messages, Crime committed by many people
6	Multiple concepts, Not enough commonality
7	Forgery, Reach of confidentiality, Corruption, Not enough commonality
8	False evidence, Multiple concepts
9	Multiple concepts
Outliers	Outliers

is not presented. Table 2 highlights the legal concepts shared in each cluster, and also shows outliers.

As mentioned in Table 2, very strong similarity of legal concepts is observed in some clusters whereas few clusters contained cases of varied legal themes.

5. **Conclusion.** Embedding using neural network framework is one of the most promising breakthroughs in the field of graph analysis. In this pilot study we demonstrated application of node2vec algorithm to obtain node embeddings for a less explored domain of legal citation network. The results obtained strengthen authors' claim of suitability of

this approach for analyzing complex citation networks. In addition to facilitating visual exploration of data, embeddings transform data into a representation which can directly be used by machine learning algorithms. Though the study limits the demonstration to data reduction and basic citation embedding-based case clustering, this approach can be easily extended to perform any complex machine learning task. We have considered subset of court cases belonging to only information technology act as the domain of law. In future we intend to expand the study to all the cases across domains in the Indian Judiciary. We also intend to consider specific characteristics of legal citation network like temporal properties, bipartite nature and acyclicity while generating embeddings to get important insights into interrelationships among various domains and subdomains of law through citation networks.

Acknowledgment. The authors would like to thank CHRIST (Deemed to be University) for its support pertaining to this study. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] D. De Felice, G. Giura and V. Verendel, Why do you quote me? Citations of Superior Court orders in Sicilian Courts, *The 2nd International Workshop "Network Analysis in Law" in Conjunction with JURIX 2014: The 27th International Conference on Legal Knowledge and Information Systems*, Krakow, Poland, 2014.

[2] K. J. Pelc, The politics of precedent in international law: A social network application, *American Political Science Association 2013 Annual Meeting Paper*, 2013.

[3] T. Neale, Citation analysis of Canadian case law, *Journal of Open Access to Law*, vol.1, no.1, 2013.

[4] F. Tarissan, Y. Panagis and U. Sadl, Selecting the cases that defined Europe: Complementary metrics for a network analysis, *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, San Francisco, United States, 2016.

[5] M. Koniaris, I. Anagnostopoulos and Y. Vassiliou, Network analysis in the legal domain: A complex model for European Union legal sources, *Journal of Complex Networks*, vol.6, no.2, 2018.

[6] R. Wagh and D. Anand, Application of citation network analysis for improved similarity index estimation of legal case documents: A study, *IEEE International Conference on Current Trends in Advanced Computing*, 2017.

[7] S. Kumar, Similarity analysis of legal judgments and applying 'paragraph-link' to find similar legal judgments, *Lecture Notes in Computer Science*, vol.7813, 2013.

[8] A. Minocha, N. Singh and A. Srivastava, Finding relevant Indian judgments using dispersion of citation network, *Proc. of the 24th International Conference on World Wide Web (WWW'15)*, Florence, Italy, 2015.

[9] A. Grover and J. Leskovec, Node2vec: Scalable feature learning for networks, *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, San Francisco, CA, USA, pp.855-864, 2016.

[10] T. Mikolov et al., Distributed representations of words and phrases and their compositionality, *Proc. of the 26th International Conference on Neural Information Processing Systems (NIPS'13)*, 2013.

[11] Y. Goldberg and O. levy, Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method, *arXiv: 1402.3722*, 2014.

[12] J. Leskovec, Christos faloutsos sampling from large graphs, *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, 2006.

[13] N. Blagus et al., Empirical comparison of network sampling: How to choose the most appropriate method?, *Physica A: Statistical Mechanics and Its Applications*, vol.477, 2017.

[14] P. Cui, X. Wang, J. Pei and W. Zhu, A survey on network embedding, *arXiv: 1711.08752*, 2017.

[15] H. Cai et al., A comprehensive survey of graph embedding: Problems, techniques and applications, *arXiv: 1709.07604*, 2018.

[16] B. Jiao et al., Comparison of biased and unbiased sampling algorithms using graph metrics, *2016 International Symposium on Computer, Consumer and Control (IS3C)*, 2016.

[17] *indiankanoon.org.*