International Journal of Engineering & Technology, 7 (3) (2018) 1286-1289

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i3.11820
Research paper

Sub graph sampling and case citation network: a case study

Nidha Khanam [1] *, Rupali Sunil Wagh [2]

[1] PG Scholar, Computer Science Department, CHRIST (Deemed to be University), Bengaluru, Karnataka, India
[2] Associate Professor, Computer Science Department, CHRIST (Deemed to be University), Bengaluru, Karnataka, India
**Corresponding author E-mail: nidha.khanam@cs.christuniversity.in*

Abstract

Graphs or networks very commonly are used to represent connected or linked data. With the penetration of www in every sphere of life networked relationships can be easily established through communication links over web and network and graph analysis come as obvious choices of data representation and analysis. There are processes which can be analysed as network not through web but through the knowledge links available in these domains. In both these cases network analysis is challenged by the enormous size of the network in terms of nodes and links. Sub graph sampling can effectively be employed on large network structures to reduce the size of data while preserving the original properties of the network. Through this paper authors present a case study on application of sub graph sampling approach to obtain reduced case citation network in legal domain.

Keywords: *Legal Sector; Network Analysis; Analytics in Legal Domain; Subgraph Sampling; Case-Citation Network.*

1. Introduction

Legal information systems or online digital legal libraries are repositories of many types of documents like acts, judgements, legal transcripts. These documents are typically knowledge bases and are used by legal professionals to solve a legal issue at hand. But since the knowledge is available in natural languages automatic extraction and interpretation by a digital system is challenging. Citations in legal domain are seen as a tool to understand knowledge transfer in the domain. Legal sector holds documents which include both structured and unstructured data. Structured data characterises citations, case id, location, date etc. Whereas unstructured data includes plain texts present in the document. The judgements in common law system delivers link to the judgements taken in similar previous cases. These links are referred to as the citations. It provides the details of judgement and ideas of the article that was considered for the judgement for a case. It also deliberates cases which fall under same category. Network Analysis is a method of signifying complicated data and structuring it in a form of network in terms of nodes and edges. It indicates node structure where similarities and patters are found using network metrics. In the previous work, authors have presented application of network structure in the analysis of Case-Citation data. Size of network in terms of number of nodes poses challenges in this analysis. This study proposes graph sampling technique to reduce the number of links based on the frequency. Sub graph sampling technique is used to obtain reduced graph from the huge network which was obtained from the case citation data. Sub graph sampling technique acts as data reduction technique for large networks and reduces the number of nodes and edges for better understanding.

2. Related work

Sampling from a network has been widely studied by the researchers. Network mining approach is used to describe about community detection, statistical properties, link prediction for large sets of data in social networks [1]. [2] Network which consists of multiple connections among any node pairs is analyzed for reducing the dimension of the network. [3] Graph Databases inherits properties of another graph which are mostly present in data structures. In this approach data manipulation are specified as type constructors and graph-oriented operations. [4] Graph which consists of several layers with multiple relationships can be shortened by considering two or three layers and by truncating the rest. [5] For sampling large and dynamic graph, simulation was done using random graphs and snapshot of the Gnutella. Both the graphs were compared which resulted in similar vertices and edges. But the degree distribution of Gnutella topology is considerably more skewed and it had more clusters. Graph was reduced by measuring Bias to check for equal probability, measured using correlation and efficiency.

Application of network analysis for legal relevant documents is associated for finding similarities based on cosine and citation similarity. Citation based similarity measure is more robust in this study. [6] Network visualization using semantic substrates is demonstrated where non-overlapping areas in which nodes are based on the node attributes. And by using users who are interactively adjust sliders to control link visibility. [7] Network having simple nodes and links which are interconnected is build up with the help of labels, directed links, link attributes and node attributes. It was proved that the network which is built based on the attribute resulted in 10-50 nodes and 20-100 links. [8] Links, relationships, nodes and judgements were applied on social network to find similarities in important cases and cases which are legally relevant, based on the statistical measures which would help the scholars and legal practitioners. Dataset gives information of arti-

Copyright © 2018 Nidha Khanam, Rupali Sunil Wagh. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

cle 264 to 300 which is a part of Indian Constitution dealing with the facts like subject of property, contracts, finance and suits. It resulted in high dispersion between the relationship by degree and structural method. [9] Topological structure is constructed for the analysis of composite data. For evolved data in legislation network temporal analysis was carried out. It resulted in evolution of legislation properties and enhanced clarification on the basis of the structure.

Analytics in legal sector offers litigators associate expansion over conflicting counsel by providing information driven insights into how judges, attorneys and parties have behaved in similar cases within the past and the way they are possible to behave within the future [10]. [11] It depends on the information of the cases which was tried in the past is related with the current under-trail cases. [12] Network sequenced transition is the tool which decreases the participation of the engineer so that they can maintain and improve their knowledge easily. [13] To solve any logic based approach like justification, argumentation in legal sector there are two difficulties they are: lack of scalability and difficulties in evaluating legal texts. [14] The structure of legal judgement is difficult where each judgement is relevant to one another. For the nominal search of similar legal documents, research is been made on web searching and information retrieval.

Citation Analysis is examined to resolve problem with web based documents including scientific papers, legal briefs, emails and Wikipedia [15]. Two main principles studies were: Appropriate region alignment and placement of regions to reduce the number of long edges in the documents. Later, author described about predator-prey relationship and US Senate voting patterns for large text documents.

Tf-idf cosine similarity measured with 2 pairs which includes formulation of term which represented the in degree and out degree for both the cases. [16] Basic work is to search and browse and also rely on the graph. It is difficult to combine semantic contents and textual link between the documents. [17] Automated mapping of legislative texts is described in this section. Query and scripts based approach was introduced on the database to extract the information; visualization was made using Hungarian legislative texts providing inner link structure.

Citation data is represented in the form of Network with network measures. To find similarities between two objects directly, analysis allows analyzing relatedness among the cases or judgements [18].

3. Methodology

The proposed work emphasizes on the network analysis of data using sub graph sampling technique by finding the frequency and by obtaining distribution of mean and variance. Figure 1 explains the methodology followed during the analysis. Various steps are described below:

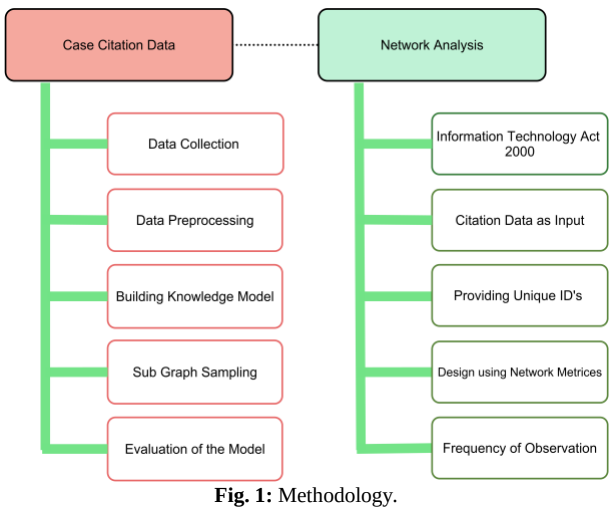


Fig. 1: Methodology.

3.1. Data collection and data preprocessing

Data used for the analysis is collected from the website "indiankanoon.org". For the purpose of this analysis authors have focused only on cases of Indian Courts. Two column format data is collected in which first column contains cases and second column contains citations. These data are semi structured and required preprocessing. Preprocessing of the dataset is done using the R tool. Generic preprocessing steps namely Sections and Act removal, Id's are given for both cases and citation.

3.2. Building knowledge model

Data represented in two columns as nodes and edges for the information obtained in preprocessing. Node represents a judgement or a case and an edge between nodes indicates citation relationship between the cases. Since edges are very important in network analysis, directed graph is then sampled for finding the frequency and visualize the graph.

3.3. Reduction of graph using sampling method

Network analysis for dense data needed to be sampled according to the exponential distribution method using simulation mean and variance which is then compared with theoretical values. Central Limit Theorem is used for obtaining graph subsample. It is the method of distribution sampling technique which reduces the sample size using sampling means and sampling variance.

3.4. Evaluation of the model

Network parameters obtained after graph sampling methods are compared with original case citation network to analyze the properties of sub-graph.

4. Results and discussion

Sub graph Sampling of Case-Citation Data– After obtaining the sub graph, frequency of mean distribution does not deviate significantly from the original network's mean value.

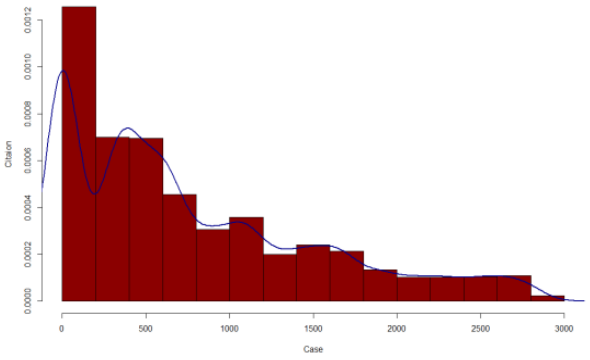


Fig. 2: Histogram for Case-Citation Data.

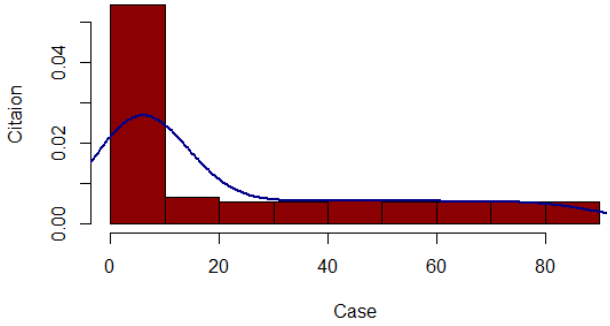


Fig. 3: Histogram with Sample Mean.

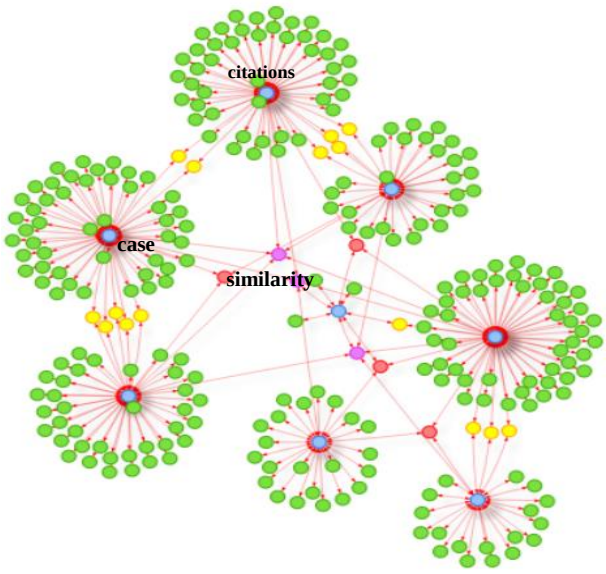


Fig. 4: Similarity Network Structure [18].

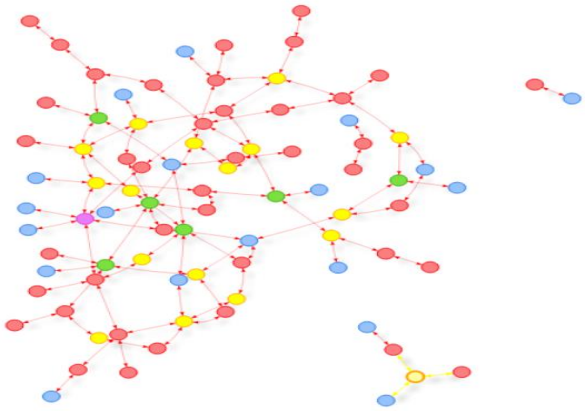


Fig. 5: Sampled Network Structure.

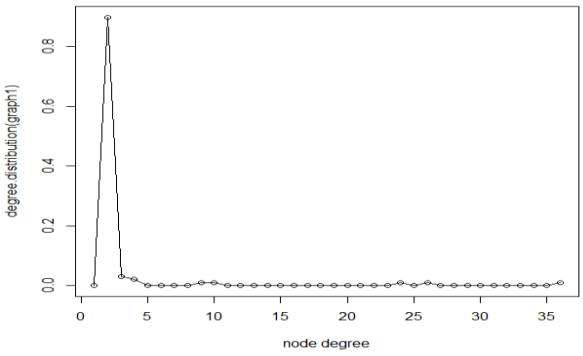


Fig. 6: Degree Distribution without Sampling.

a) Case Citation Data

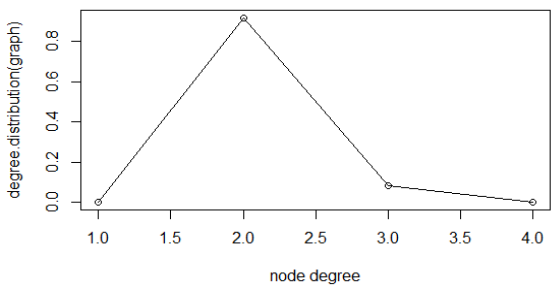


Fig. 7: Degree Distribution with Sub Graph Sampling.

b) Case Citation Data using Central Limit Theorem

Table 1: Network Metrics Before and After Sampling Techniques with Fig. 6 and Fig. 7

Property	Results	Property	Results
Length	10	Length	4
Vertices	283	Vertices	154
Edges	310	Edges	166
Density	0.00388442	Density	0.0005862682
Diameter	1	Diameter	3
DegreeDistribution	0.89	DegreeDistribution	0.91

As it is evident from the table, and the degree distribution graph, the characteristics of original graph are restored even after application of graph sampling;This implies that the reduction in nodes and edges has a very negligible effect on the generic characteristics of the original network.

5. Conclusion and future work

Finding related cases is one of the most researched problems in legal domain. Due to large data, plotting a network is non-understanable. So sub-graph sampling technique is prepared for analysis which shows reduced number of graph with nodes and edges. Through this paper authors have presented sub graph sampling for case citation network of Indian court judgement which can be used not only to reduce the complexity of data but also to visualize it in a precise way. Information can be analyzed using the judgement year, courts and judges for most relevant laws. If edges can weighed by designing knowledge based weights, such structure can be very useful in understanding legal knowledge. Other sampling methods can further be explored for finding most appropriate approach for case citation network.

References

[1] Ana Paula Appel, Luis G. Moyano, "Link and Graph Mining in the Big Data Era", *Springe*, 2017. https://doi.org/10.1007/978-3-319-49340-4_17.
[2] Michele Berlingerio, Michele Coscia, FoscaGiannotti, Anna Monreale, Dino Pedreschi "Multidimensional networks: foundations of structural analysis", *Springer*, 2011.
[3] Rohit kumar Kaliyar, "Graph Databases", *ICCCA*, 2015.
[4] Robert Bredereck, Christian Komusiewicz, Stefan Kratsch, HendrikMolter, Rolf Niedermeier and Manuel Sorge, "Assessing the Computational Complexity of Multi-Layer Subgraph Detection", *Springer*, 2016.
[5] Daniel Stutzbach, Reza Rejaie, Nick Duffield, Subhabrata Sen, Walter Willinger, "Sampling Techniques for Large, Dynamic Graphs", IEEE, 2007.
[6] Rupali Sunil Wagh, "Application of Citation Network Analysis for Improved Similarity Index Estimation on Legal Case Documents: A study", *ICCTAC*, 2017.
[7] Ben Shneiderman, AleksAris, "Network Visualization by Semantic Substrates", *IEEE*, 2006.
[8] John Stasko, CarstenGörg, Zhicheng Liu, "Supporting investigative analysis through interactive visualization", *IEEE*, 2008.
[9] MariosKoniaris, LoannisAnagnostopoulos, YannisVassiliou, "Network Analysis in the Legal Domain: A complex model for European Union legal sources", *Journal of Computer Networks*, 2015.
[10] Dan Steiner, "Data Analytics and Your Law Firm, Law Technology Today", *Law and Innovation Conference*, 2016.
[11] Owen Byrd, Legal, "Analytics vs. Legal Research: What's the Difference? Law Technology Today", *Law and Innovation Conference*, 2017.
[12] Andrew Stranieri, John Zeleznikow, "Tools for Intelligent Decision Support System Development in the legal domain", *IEEE*, 2000.
[13] Karl Branting, "Data-centric and logic-based models for automated legal problem solving", *Springer*, 2017.
[14] Sushanta Kumar, P. Krishna Reddy, V. BalakistaReddy, Malti Suri2, "Finding Similar Legal Judgements Under Common Law System", *Springer*, 2013.
[15] Pak Chung Wong, Chaomei Chen, CarstenGörg, Ben Shneiderman, John Stasko, Jim Thomas, "Graph Analytics—Lessons Learned and Challenges Ahead", *IEEE*, 2012.

[16] AkshayMinocha, Navjyoti Singh, ArjitSrivastava, "Finding relevant Indian judgemegnts using Dispersion of Citation Network", *ACM,* 2015.
[17] GáborHamp, RékaMarkovich, "Automated Reference Extraction in Hungarian Legislative Texts and Visualization of their Inner Link Structures", *Openlaws Open Data Workshop, NAIL,* 2015.
[18] Nidha Khanam, Rupali Sunil Wagh, "Application of Network Analysis for Finding Relatedness among Legal Documents by Using Case Citation Data", Journal *on Information Technology, UGC,* 2017.