Application of Citation Network Analysis for Improved Similarity Index Estimation of Legal Case Documents : A study

Rupali Wagh
Research Scholar Jain University
Bangalore, India
Rupali.wagh@christuniversity.in

Deepa Anand
CMR Institute of Technology
Bangalore, India
Deepa.a@cmrit.ac.in

***Abstract*—With the availability of information in online databases, information retrieval has become back bone of many processes today. Significant share of this online data comprises of unstructured and textual data. There are many domains like legal domain which rely solely on information stored in various legal documents. Legal domain is considered to be very complex and its processes are largely dependent on knowledge interpretation by human expert. Establishing relevance and similarity between two cases based on expositions in various legal documents is the most common but non trivial task performed by a legal expert. This paper discusses application of network analysis to compare two approaches for finding legal document similarity – a) Cosine similarity b) Citation based similarity. Results show that citation based similarity measure is more robust in determining parallel among cases**

Keywords—Document similarity, Legal documents, citation network, tf-idf scoring, cosine similarity

I. INTRODUCTION

Legal domain generates huge information in the form of text and documents. Legal information can be categorized under various headings like court transcripts, verdicts, statements and affidavits etc. Such documents are repositories of useful information regarding the interpretations of law and a legal researcher has to study such documents. Legal informatics is a specialized field which aims at application of information technology for effective management of legal information. Researchers, both from information technology and law have been contributing to provide effective knowledge management in the domain of law. In last few years, we have witnessed progression from information processing to knowledge management across domains. Evolving analytics is contributing to more and more intelligent solutions facilitating data driven business decisions. Legal knowledge is inherently complex and is stored in legal documents using natural language which makes it more complex for automatic analysis and interpretation. For any legal professional, extracting relevant precedents for defining the context of the problem at hand is very crucial. Under common law system, a legal problem is studied as a context which can be reconnoitered with the help of precedents, thereby making process of legal reasoning and decision making heavily dependent on information stored in text documents. Many online legal databases provide easy access to such legal documents. These databases allow users to search based on legal terms like act, court, judgments, orders etc. These search options require the query to be very precisely stated with terms very specific to the domain. Though free text search options are available for non-expert users, the performance of search queries still leaves much to be desired. Due to the complex nature of legal knowledge present in legal document, the efficacy of traditional document similarity models like vector space model is limited. Citations made by a legal document play a very important role in deciding the parallel among various cases. This paper compares citation based document similarity measures with cosine similarity measure for legal judgments. The results are analyzed using network metrics. It is further shown that the citation based similarity is closer to human expert rating on similarity of legal documents.

II. RELATED WORK

Documentary informatics using Natural language processing (NLP), Artificial intelligence, Machine learning and network analysis are major knowledge management approaches used for legal domain. Since legal knowledge in various legal documents is written in natural languages, NLP based applications are used appropriately in this domain. Previous studies show that Named entity recognition and POS tagging can be effectively used for legal entity extraction, metadata extraction and identification of actors etc. [1, 2, 3]. Wide ranging machine learning techniques like clustering, and text classification, as in any other information retrieval system, can be used for grouping of similar documents for improving performance [4, 5, 6]. Network analysis is used to study and analyze collection of documents which are interrelated. When represented as a network, a document in the collection can be modeled as a node whereas an edge represents relationship between two documents. In the era of digital libraries and with the document collection size growing exponentially, finding related documents is becoming increasingly difficult. Network analysis can be used aptly particularly in domain specific information retrieval systems for enhancement of

978-1-5090-4997-4/17/$31.00 ©2017 IEEE

retrieval results. In legal domain, network analysis is applied to wide ranging problems from finding complexity of a legal document, designing conceptual maps for complex legal document to legal recommender systems [7]. Social network analysis is providing a new dimension to analyze criminal networks and providing important knowledge. An edge of any network has several attributes having greater significance in the analysis of network. Based on the domain and the concept being analyzed as network, many approaches of defining links have been proposed and used. Information stored in the document can be used to decide upon the weight of an edge [8]. Most commonly used approach is finding the similarity between two documents using a text analysis model to use it as weight of the edge. Application of text similarity measure like cosine similarity and vector space models suffer from the problems of synonyms and homonyms. Additionally, entire text in the document may not be an appropriate indicator of similarity between documents. Using different weight value for different sections of the document for calculation of overall weight of the link can be used in case of documents following similar structure [9]. Advanced text analysis can be used to identify and extract important paragraphs in the documents based on domain knowledge

Citation network is a very important tool to study knowledge dissemination in all domains. Most basic approaches treat citation of a document in another document as a link or an edge. Research articles are the powerful source of information which presents new knowledge essentially as an extension to the existing one. Citation networks are extensively used to understand and analyze various aspects of knowledge transfer. Research paper citation networks can also be studied to understand other associations like affiliations and citation [10]. Patent analysis is a field closely related with research and is heavily impacted by previous knowledge. Citation analysis is used to get more insights into related previous patents and their interrelationships [11]. Application of query specific citation network for expansion of query for patent search has also been suggested [12]. A case citation is one of the primary knowledge components and is used widely by legal experts for deciding the relevance of precedent cases to the context being studied. There is a continuous process of new knowledge creation based on prior knowledge in the form of new verdicts and judgments. Application of citation network analysis for extracting important information from Canadian case law corpus is discussed in [13]. Simple network metrics viz. degree distribution can reveal significant information like citation patterns and time period analysis of cases. Significance of citation analysis in improving the performance of information retrieval systems is discussed in [14]. The importance of the cited case can be decided based upon the perspective in which it is used. To exploit this information ranking method based on the paragraphs in which citation appears are also proposed for Indian Supreme Court Judgments [15]. Though case law predominantly relies on precedents, appropriate weight for legislation references is required for better understanding of relevance. This paper considers acts cited by different judgments for calculation of edge weight. It further compares this network with that of obtained by using cosine similarity measure as edge weight. The results are then compared with human expert's rating to show that citation based similarity methods outperform the traditional cosine similarity approach.

III. METHODOLOGY

This work is comparative study of two weight measures for calculating similarity among legal judgments. As highlighted in the previous sections, network analysis is used for comparison of results with human expert score. Figure 1 explains the methodology followed during the course of the study. Various steps followed are described below.

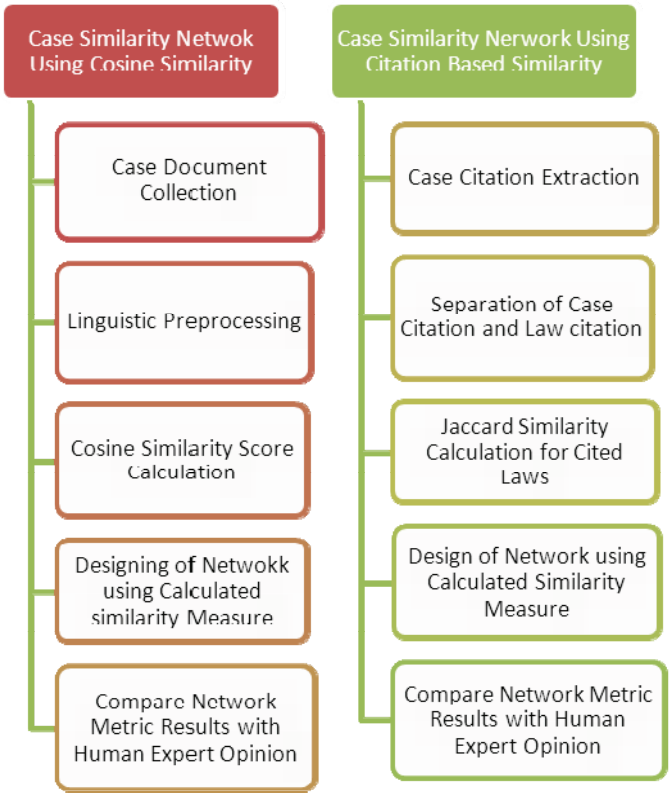

Figure 1 - Methodology

1. **Data Collection and data preprocessing** – Court judgments based on a legal concept called as "resjudicata" forms the data corpus for this study. Resjudicata refers to a legal matter that is already judged. This data is available online on the website "indiankanoon.org". These documents are totally unstructured and require linguistic preprocessing to transform unstructured data into suitable structured information. Generic preprocessing steps namely stop word removal, Punctuation symbol removal; stemming and Normalization are applied on the case judgment document. Term document matrix is then constructed using widely used tf-idf scoring. Citation data for every case is recorded separately.
2. **Designing case similarity network** – This step focuses on exemplification of legal documents as network with

appropriate representation of nodes and edges for the information obtained in preprocessing. A node in this study represents a judgment or a case and an edge between nodes indicate that the two cases are related. Since edges are very important in citation network analysis, weights should be assigned to edges using suitable measures. These weights can act as relevance measures for two cases. The study has used two different weight measures namely cosine similarity for the entire document and jaccard similarity for citations in the document.

a) Cosine similarity for documents represented using vector space model has been always been very popular due to its simplicity. Cosine similarity of two length normalized vectors (documents) d1 and d2 is inner product of defined by the formula

$$Sim(d1,d2) = Cos(d1,d2) = d1.d2$$

Matrix of cosine similarity values for case documents based on term document matrix obtained in step 1 is derived. These similarity values are used as edge weights while designing case network.

b) Citations extracted from case judgments are treated as sets to obtain Jaccard similarity matrix Jaccard similarity measure is applied on set representation of citation data. For two sets c1 and c2, Jaccard similarity is defined by the formula

$$Sim\ (C_1, C_2) = |C_1 \cap C_2| / |C_1 \cup C_2|$$

These similarity values are used as edge weights while designing case network.

3. **Evaluation of the knowledge model** – Both Network models obtained in the previous step are then analyzed with the help of network metrics like degree distribution, centrality, and connected components. Instead of considering only one to one link as the only indicator of similarity, existence of a path from one node to other node is also considered for deciding similarity. Hence connectedness of cases is also analyzed using the network representation.

IV. RESULTS AND DISCUSSION

The primary goal of network analysis in this study is to analyze similarity index in the collection of documents. Various network algorithms and metrics like centrality, betweenness and connectedness can be used to analyze interrelationships, this work mainly focuses on edges and degree distribution. An edge represents connection between two nodes (court judgement this study) of a network. Weight attribute of an edge can reveal about the strength of the association between two documents which can be used to rank the similarities among the documents in the collection.

In all figures and tables X_i represents a legal case. Figure 2 and Figure 3 represent network constructed using cosine similarity respectively without any filter and with filter on edge weights. Whereas figure 4 represents network constructed using citation similarity as mentioned in the methodology.

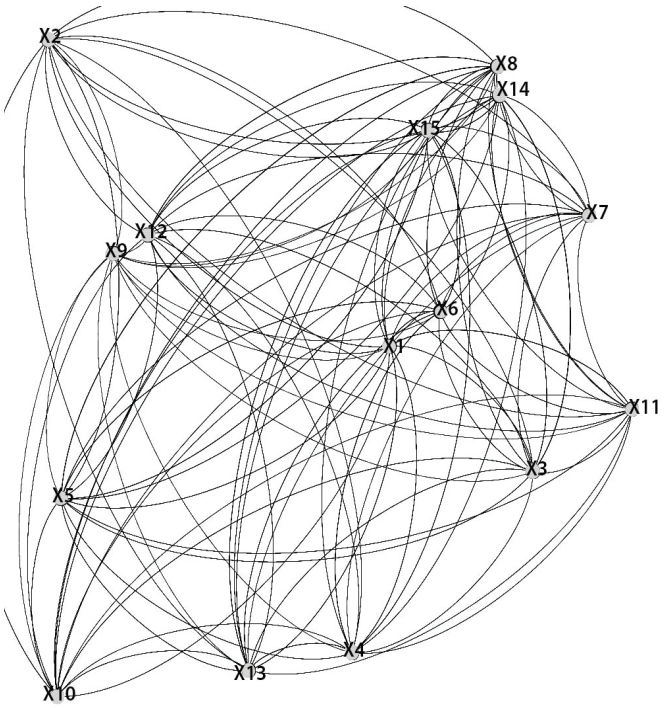

Figure 2– Cosine Similarity Network without Filter on Edge weight: All nodes treated as similar

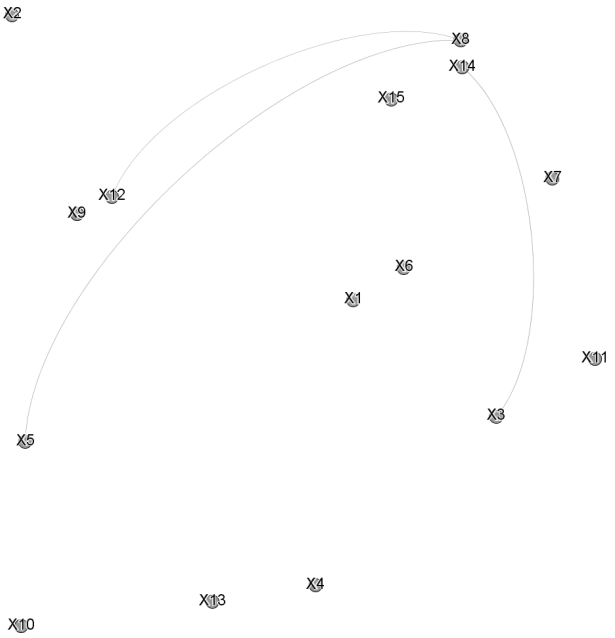

Figure 3 – Cosine Similarity Network with Filter on Edge weight: Isolation of few nodes

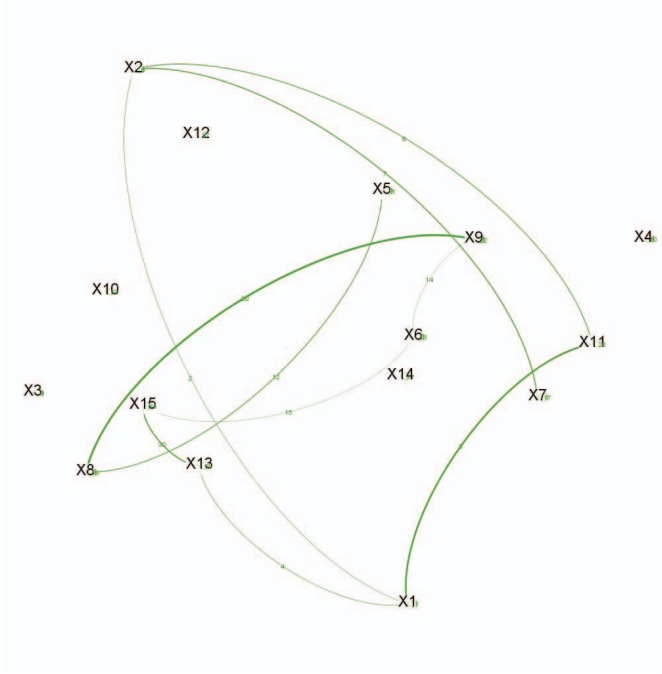

Figure 4 – Jaccard Network without Filter on Edge weight

In the networks obtained, presence of a direct link or a path from one node to other indicates similarity. It is evident from the visual representation of the network that the cosine similarity of documents has limited ability to provide insights into similarity with other documents in the collection. The network exhibits non-or-all property, if no filter on edge weight (figure 2) is applied all the cases are treated as similar but with the application of filter on edge weight, some of the nodes (figure 3) become isolated. In contrast, as shown in the figure, network constructed based on citation based similarity demonstrates balanced link patterns among documents and is more informative in understanding the similarity of cases. **Table 1** summarizes the comparison of obtained similarity with human expert opinion. It is evident from the table that the citation based similarity results are more close to human expert's opinion regarding similarity of cases. Cosine similarity, which is based on terms present in the judgement document considers all cases similar when edge weight is not applied. If edge weight is applied the network gets divided in disjoint groups.

Table 1 : Similarity Comparison with Human Expert

Similar cases According to human expert	Similar cases Using Citation Similarity	Similar Cases using Cosine similarity
X1, X2, X6, X7, X8, X9, X10, X11, X12, X13, X14, X15	X1, X2, X5, X7, X8, X9, X11, X13, X14, X15	X5,X8,X12 – Ist Group X3 and X14 – IInd Group

Connectedness is a measure of how much a node is linked with other nodes in the network and is valuable in understanding information dissemination. Values obtained for most basic connectedness parameters are given in **Table2**. These values also indicate that citation based similarity measure can be used effectively to explore reachability of a case from another and to establish equivalence among cases in legal research. In cosine similarity network every node (case) is connected (similar) to every other node hence high average degree value and only single large connected component id observed. But in citation based similarity though degree value is less, there exit paths among node that can be taken as indicator for relevance among connected cases.

Table 2 : Network Metrics

Type of Network	Avg Degree	Avg Path Length	Connected Components
Cosine Similarity Based	14	1	1
Citation Similarity Based	3.3	3.35	3

V. Conclusion and Future Work

Finding similar cases is one of the most researched problems in legal informatics. This paper shows that application of citation network analysis can be effectively used not only for similarity index estimation but also for understanding interrelationship among various legal concepts through citation links. A network can further be analyzed by applying

link analysis algorithms. Page rank and HITS are two commonly used algorithms for network link analysis. Though these algorithms are primarily used for web pages over internet, they can be applied for many generic networks by drawing analogical similarities. Utilizing additional information like year of judgments, courts and judges, this work can further be extended to get more insights on most authoritative judgments, time period analysis of a case, most relevant and irrelevant laws and decisions.

REFERENCES

1. Monica Palmirani and Raffaella Brighi, Metadata for the Legal Domain, Proceedings of the 14th International Workshop on Database and Expert Systems Applications (DEXA'03), 2003, IEEE
2. Mihai Surdeanu, Ramesh Nallapati and Christopher Manning, Legal Claim Identification: Information Extraction with Hierarchically Labeled Data, Stanford University
3. GUIRAUDE LAME, Using NLP techniques to identify legal ontology, Artificial Intelligence and Law (2004) 12: 379–396, 2006, Springer
4. Biao Fan, Tao Liu, He Hu and Xiaoyong , "*Law Text Clustering based on Referential Relations",* 2010 IEEE, 978-0-7695-4106-8/10, DOI 10.1109/ChinaGrid.2010.22
5. Qiang Lu, William Keenan, Jack G. Conrad, Khalid Al-Kofahi, "*Legal Document Clustering with built in Topic Segmentation*", CIKM'11, October 24–28, ACM 978-1-4503-0717
6. Dozier Christopher and Jackson Peter, **"***Mining text for expert witnesses",* 2005 IEEE
7. Pre Proceedings of 2nd International Workshop on Network Analysis in Law held on December 10, 2014 in Krakow in conjunction with the 27th JURIX conference on Legal Knowledge and Information Systems.
8. Chih-Hung Hsieh, Louis Y. Y. Lu, John S. Liu2, Alexander Kondrashov, A Literature Review with Citation Analysis of Technology Transfer, 2014 Proceedings of PICMET '14: Infrastructure and Service Integration
9. Weiling Chen and Gang Wang and Fengxia Yin, Document Similarity Calculation Model of CSLN,2014 IEEE
10. Hiran H. Lathabai, Thara Prabhakaran, Manoj Changat, Affiliations network analysis in scientific citations: A Case study of Information Technology for Engineering, 2014 International Conference on Data Science & Engineering (ICDSE)
11. M. Karvonen, T. Kässi ,Patent Citation Analysis as a Tool for Analysing Industry Convergence, 978-1-890843-23-6/11
12. Parvaz Mahdabi, Fabio Crestani, The effect of citation analysis on query expansion for patent retrieval, Information Retrieval, Volume 17, Issue 5-6 , pp 412-429, Springer
13. Thom Neale, Citation Analysis of Canadian Case Law, 2011
14. Anton Geist , The Open Revolution: Using Citation Analysis to Improve Legal Text Retrieval, European Journal of Legal Studies, < http://www.ejls.eu/6/81UK.html>
15. Sushanta Kumar, P. Krishna Reddy, V. Balakista Reddy, Malti Suri, Similarity Analysis of Legal Judgments and applying 'Paragraph-link' to Find Similar Legal Judgments, Databases in Networked Information Systems, Lecture Notes in Computer Science Volume 7813, 2013, pp 103-116