International Journal of Engineering & Technology, 7 (2) (2018) 855-858

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i2.9657
Research paper

Similarity analysis of court judgments using clustering of case citation data: a study

Disna Davis Kachappilly [1] *, Rupali Sunil Wagh [2]

[1] PG Scholar, Computer Science Department, Christ University, Bengaluru, Karnataka, India
[2] Associate Professor, Computer Science Department, Christ University, Bengaluru, Karnataka, India
**Corresponding author E-mail: disna.kachappilly@cs.christuniversity.in*

Abstract

Information retrieval (IR) is an automatic mechanism to extract required information from a collection of unstructured or semi-structured data. IR systems minimize the effort of a user to locate the information based on the requirements. Clustering of documents is carried out as a preprocessing step for filtering irrelevant information in an IR system. Legal domain is a producer as well as consumer of huge information which also contains invaluable legal knowledge and its interpretation. Knowledge based legal information retrieval systems is need of the day. Citation analysis is a technique to find the hidden relationships between the documents and is used for understanding knowledge transfer across various domains and hence becomes very important in legal domain. In this study, similarities among documents are analyzed using data clustering when applied on data of citations in court judgments.

Keywords: *Citation Analysis; Data Clustering; Information Retrieval; Legal Domain; Similarity Search.*

1. Introduction

Legal domain through its various processes generates a large set of information in the form of legal documents and texts. These legal documents are classified under various headings such as court verdicts, statements, FIR etc. One important category of such documents is collection of judgments belonging to different courts which are given by judges. These documents contain useful information for legal researchers to study about a particular case. Legal informatics is a field which aims at providing effective management of these legal documents. In order to study a case, legal professional needs to browse through the vast collection of documents stored in the database to find the similar cases which is a time consuming task. Knowledge of these legal documents is usually stored in the form of natural language which makes it difficult for analysis and interpretation. With the advent in internet search technologies, several efficient online legal databases facilitate this search process for user. But this search is predominantly keyword based. In order to retrieve a precedent case, the user or researcher has to specify the query using terms which are specific to the domain and thus demands sound domain knowledge. Due to this complex nature, efficiency of similarity search operation is limited.

Legal domain produces documents which include both unstructured data and structured data. Plain texts present in the document represents unstructured data whereas citations, case id, date, location etc. which are present in the document can be considered as structured information. Citations in a legal document plays a very important role in showing basic similarity

between various judgments. It conveys the information that the particular judgment is referring to another judgment and hence the two documents share similar legal concept. Citation analysis is used in legal domain to build case and citation network. This paper proposes use of citation data for finding similarity among cases using clustering techniques. Analysis is performed on court judgments in India related to cyber-crimes which are covered under Information Technology Act 2000 in the Indian Constitution. Cyber-crimes cover crimes such as e-mail fraud, identity theft, spam, illegal downloading etc. which are committed via internet and digital technologies. The information technology act 2000 was introduced by the Parliament of India to provide a control over the cyber related crimes by introducing a legal framework.

Clustering is a method used to group the data based on inherent similarity. In this paper, clustering is used to group the interrelated documents based on their citations. Clustering is often referred to as a first step in data mining for determining groups containing similar objects. It is an unsupervised learning technique which identifies groups of related records that can be used as a starting point for exploring further relationships. In this proposed work a case i.e. a judgment is considered as an object and citations referred in the judgment are considered as its attributes. Thus if citations are similar it can be concluded that the cases are similar. Application of clustering algorithm hence can provide a group of similar cases based on the similarity in citations.

2. Analytics in legal domain

Legal analytics [1] plays a very important role in the legal domain and encompasses the techniques of extracting valuable information from the data which are present in a case or a legal document. Machine learning techniques also suggested in legal analytics to process, analyze and structure the raw data present in a case document. Artificial intelligence approaches [2] for analyzing data is changing the way a lawyer analyzes a case. Artificial intelligence is a technique which is used to learn how to complete a specific task that is usually done by the human. There are several tools developed for systems based on the law domain for analyz-

Copyright © 2018 Disna Davis Kachappilly, 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.

ing the data. One of such early tools introduced is used to categorize all the legal tasks according to the class labels on computer based implementations [3]. The sequenced transition network is also a work in similar direction which caters toward decreasing the need for prior knowledge about legal data. Approaches based on logic proposed in [4] to solve a legal problem highlight two main problems: creating techniques for representations of the legal text and the difficulties involved for evaluating the legal terms mainly based on languages. The nature of legal rules or judgments is very complex and demonstrates citations based relationship with other judgments [5]. As suggested by authors, this citation links based method for finding case similarity has been proved to be more accurate. This approach can also be coupled with content based case judgment similarity for complete similarity estimation can be made more efficient by using an approach called paragraph-link approach.

Citation analysis is a very important approach which is used to study knowledge transfer in many domains. Interrelation among the patent and research publications can be studied with help of citation analysis. Citation from patents [6] to the publications provides useful proofs related to the academic research. Since the patent citation is of large-scale, evaluations required for the patent and publication links is done by the help of automatic searchable engine. Methods such as bibliometric are used to evaluate the publications based on their citations such as Scopus and Google. Since Google patents do not provide indexing for the academic citations, the process of bibliometric study on patent citation is time consuming. This is very difficult for the evaluators to get the overall impact of the large number of legal articles. The scientific references present in patent documents represent the interaction between science and technology. [7] presents a statistical review of such interaction using citation analysis. Paper indicates that the motivation of science is about half of the entire inventions. About 30% of patents which are motivated by science does not contain logical references. Additionally, 20% of cited logical references is assessed as unimportant by the inventor. To review the importance of analysis in patent citation, the characteristics of patent citation with respect to logical literature should be identified. Novel mapping to recognize technology-relevant research [8] is introduced on papers referring to SNPRs. Citation analysis is back bone of legal domain. It is difficult for legal researchers and analyst to evaluate the impact of large number of legal articles. To overcome this problem, an approach such as Bing searches were introduced which is combined with the filtering of results automatically for the duplicate data produced by the Bing search [6].

Similar precedent extraction is most important but non-trivial process in legal domain. Many solutions are proposed for this said problem which are predominantly text and based [9] mainly discusses about analysis of legal document that includes text information in the form of semi structured and unstructured form. Natural languages are used to store this legal text information in the database. For extracting the text data from the database text mining techniques are used in this work. This paper aims in using the text mining techniques for grouping of the legal text documents based on its contents without considering the external query input. [10] discuses about the main issues regarding the processing of information in legal domain which includes issues such as handling the complexity of knowledge in the legal domain by filtering the techniques and methods and identifying the ways to store and reclaim the required data.

Use of citation data for finding interrelated information is explored by many researchers in various domains. Combining Citation data analysis along with clustering is one of the very popular approaches experimented by researchers in information retrieval domain. [11] demonstrates the use of two software tools such as CitNetExplorer and VOSviewer which can be used to cluster or group the citation data or publications and then each cluster can be analyzed. In this paper clusters are made based on the direct relation of their citation. Both tools can be used to analyze the cluster solution. CitNetExplorer is used to make the groups or clusters of the publications based on the direct relation of their citations and can also be used to visualize the citation clusters, where publications are shown in the time axis with different colors that represent each cluster. One of the major functionality of CitNetExplorer is to analyze the clustering solutions in different level. Searching for a particular paper can be done using title, author name etc. CitNetExplorer is mainly used to analyze the solution of the cluster in different level whereas VOSviewer which is one of the components of CitNetExplorer, can be used to analyze a cluster solution at an aggregate level. [12] discuss about a new approach of grouping the documents based on citation context data. With this, features of a specific citation word are compared with the textual representation of the original document. Comparison between the link based clustering algorithm with the text based clustering algorithm is discussed in this paper [12] which identifies the similar documents based on the citations which is grouped under some categories.

Data clustering is one of the most basic unsupervised techniques used to group similar items. K means clustering algorithm is very popular due to its simplicity. Many variations to basic k means algorithm are suggested in literature. [13] discusses about the novel binary search algorithm for clustering the data to find the quality of each cluster and also produces the same solution each time the algorithm is performed on the dataset. [14] proposed an algorithm called the psFCM algorithm which is the modified version of fuzzy C-Means algorithm. This algorithm reduces the computational time which is required to divide the dataset into the required number of clusters. The psFCM algorithm is divided into two phases. Phase I the dataset is divided into small blocks using the k-d tree method [15] so that the original data set is reduced into simplified data set with some unit of blocks. The clusters are calculated by calculating the cluster center. All the patterns present in the blocks are replaced by the center value that is calculated before. As a result, a huge number of patterns present in the original data set are rapidly reduced into smaller number of centroid block datasets. With the simplified dataset we find the actual cluster center using the fuzzy C means algorithm. In phase II, with the cluster centers found using the fuzzy C means algorithm is initiated with the final clusters that was already found in phase I.

Literature suggests that legal documents due to their complexity pose many challenges in knowledge extraction. Though legal information can be processed as unstructured text, legal citations are very important in understanding relationships among judgments passed by courts. And hence legal issues discussed in cases. This study proposes the application of basic k means clustering algorithm on case citation data to analyze similarities among cases.

3. Proposed methodology

The below figure, discusses steps of the proposed methodology in the study:

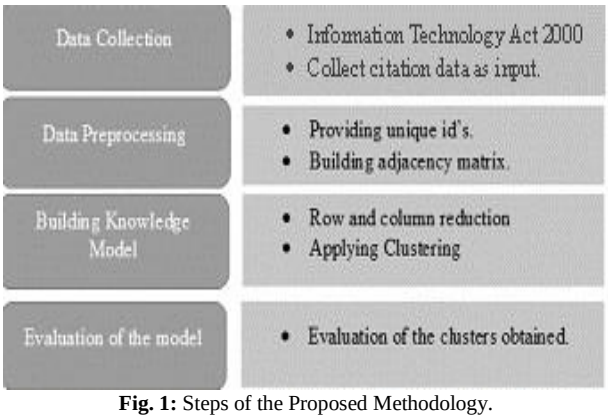


Fig. 1: Steps of the Proposed Methodology.

3.1. Data collection

The data required for this process is obtained from freely available online legal database "Indian Kanoon.org" website. For the pur-

pose of this study authors have focused only on cases under information technology act, of Indian constitution. The data is collected in two column format in which the first column contains the cases and the second column contains its citations.

3.2. Data preprocessing

The preprocessing of the dataset is done using R tool. The semi-structured data is pre-processed using the following steps: First, unique IDs are given to the cases as well for its citations such as the cases which are repeated will have the same id's. Second, the two column dataset is converted a row column format such that we will have multiple columns with the citations. Third, the dataset is converted into binary matrix such that the cases which are cited will have the value 1 or else will have the value 0.

3.3. Building knowledge model

After preprocessing, the dataset is reduced to 583 observations and 2271 variable. To reduce the dimensionality, further filtering on number of citation is applied. Clustering commonly refers to grouping the objects based on some similarity. The study is done using k-means clustering in which the number of clusters required is specified. Distance between each data point to the centroid value is calculated and then the data point put into the respective clusters. Basically, k-means clustering method is used to partition the entire dataset into k groups [16] and then selects the respective cluster center and makes use distance formula to calculate the distance between the data. The main issue of
k means clustering is to choose the value of k. In this study, in order to choose the optimal number clusters, a method called the elbow method is used. K means clustering is then applied on the preprocessed data.

4. Results and discussion

Clusters thus obtained by applying the algorithm are evaluated for its performance. SS stands for sum of squared distance. Total_SS is derived when the sum of squared distance of each data point is computed with the centroid value. Instead of computing the global mean, we compute the mean of each group (here, two groups are there) and then multiply the squared distance of each mean to the global mean from which we derive between_SS. The ratio of between_SS and Total_SS gives us 90.2% (0.902) which indicates a good fit of K-means clustering.
The below figure explains about the results of elbow method used in order to assess the optimum value of k.

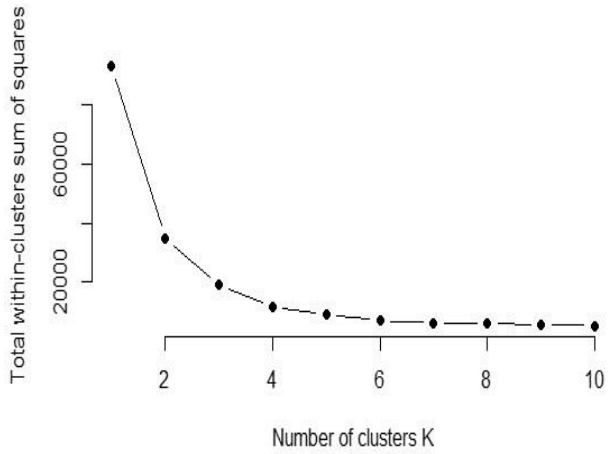


Fig. 2: Finding Optimum Number of K Using Elbow Method.

Based on the results of elbow method K-means clustering algorithm was applied in the data set for k= 3, 4, 5. The below table explains about the performance that are obtained for the k-means clustering algorithm performed for three, four, five clusters.

Table 1: Performance of K-Means Clustering

Clusters	between_SS / total_SS	Cluster sizes
K=3	79.2%	379, 14, 67
K=4	87.5%	26, 75, 4, 355
K=5	90.2%	4, 53, 21, 286, 96

As seen in the table the ratio of sum of squares between clusters with total sum of squares for obtained cluster confirm the validity of groups extracted after analysis.
To analyze the quality of individual clusters, silhouette values for individual values were plotted as shown in the figures.

When k=3

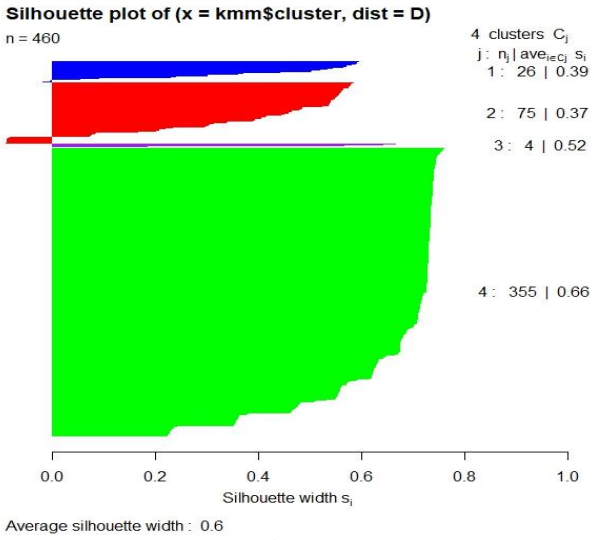


Fig. 3: Results of K-Means When K=3.

When k=4

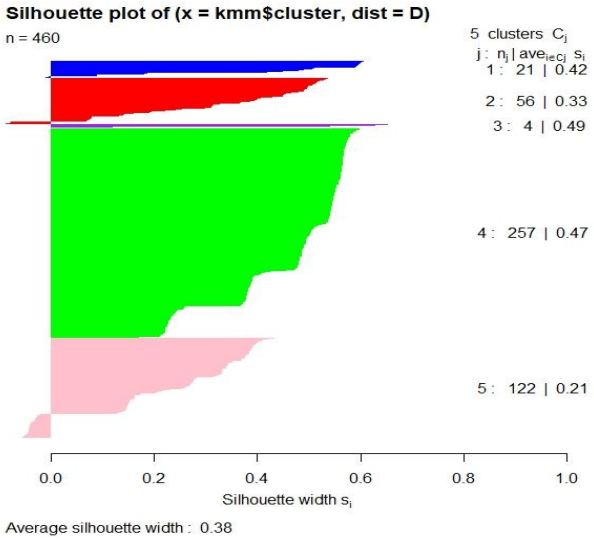


Fig. 4: Results of K-Means When K=4.

When k=5

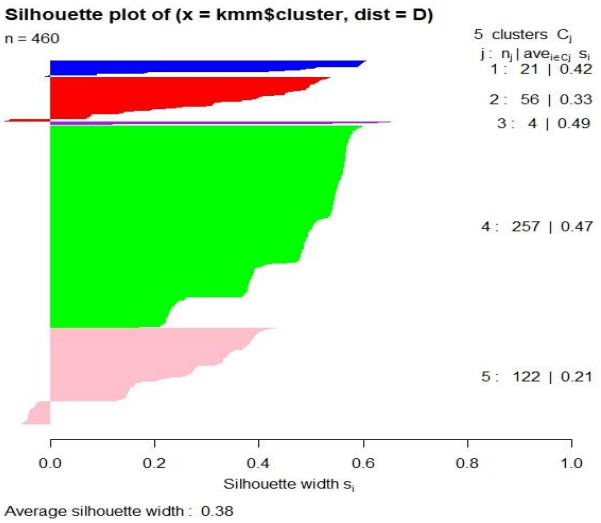


Fig. 5: Results of K-Means when K=5.

Silhouette plot indicates that as number of cluster k is increased silhouette values tend to get negative. Best result is shown for value of k=3. While clusters with maximum and minimum size are less affected by values of k, moderately sized cluster tends to show decline in the values. This suggests that the data may contain overlapping groups which can be more effectively analyzed using fuzzy and rough set approaches.

5. Conclusion

Through this study authors have presented their efforts towards reducing manual overload of legal professionals in finding similar judgments. Results obtained during study justify the approach followed though the results can be improved further by using variations in basic k-means algorithm. Such citation based clustering of data for similarity analysis can be used for information retrieval in legal databases. Sparse citations in legal judgments poses challenges regarding the efficacy of the approach, but if citation data is augmented with specific legal knowledge, application of clustering may result into better grouping.

References

[1] Owen Byrd, Legal Analytics vs. Legal Research: What's the Difference? Law Technology Today, 2017.
[2] Julie Sobowale, How artificial intelligence is transforming the legal profession, ABA Journal, 2016.
[3] Andrew Stranieri, John Zeleznikow, Tools for Intelligent Decision Support System Development in the legal domain, 2000.
[4] Karl Branting, Data-centric and logic-based models for automated legal problem solving, Springer, 2017.
[5] Sushanta Kumar, P. Krishna Reddy, V. Balakista Reddy, Malti Suri2, Finding Similar Legal Judgements under Common Law System, 2013.
[6] Kayvan Kousha, Mike Thelwall, Patent Citation Analysis with Google, Association for Information Science and Technology, 2015.
[7] Julie Callaert, Maikel Pellens, Bart Van Looy, Sources of Inspiration? Making Sense of Scientific References in Patents, 2014.
[8] Anthony F.J. van Raan, Patent Citations Analysis and Its Value in Research Evaluation: A Review and a New Approach to Map Technology-relevant Research, Journal of Data and Information Science, Vol. 2, 2017.
[9] Rupali Sunil Wagh, Knowledge Discovery from Legal Documents Dataset using Text Mining Techniques, International Journal of Computer Applications, Volume 66, 2013
[10] Rupali Sunil Wagh, Exploratory Analysis of Legal Documents using Unsupervised Text Mining Techniques", International Journal of Engineering Research & Technology, Vol.3, 2014.
[11] Nees Jan van Eck, Ludo Waltman, Citation-based clustering of publications using CitNetExplorer and VOSviewer, 2017.
[12] Bader Aljaber, Nicola Stokes, James Bailey, Jian Pei, Document clustering of scientific texts using citation contexts, 2009.
[13] Abdolreza Hatamlou, In search of optimal centroids on data clustering using a binary search algorithm, 2012.
[14] Ming-Chuan Hung and Don-Lin Yang, an Efficient Fuzzy C-Means Clustering Algorithm, 2001.
[15] J.L. Bentley, Multidimensional binary search trees used for associative searching, Communications of the ACM, Vol. 18(9), 1975, 509-517.
[16] Kiri Wagstaf, Claire Cardie, Constrained K-means Clustering with Background Knowledge, Proceedings of the Eighteenth International Conference on Machine Learning, 2001,577–584.