International Journal of Engineering Research and Technology.
ISSN 0974-3154 Volume 11, Number 3 (2018), pp. 373-381

http://www.irphouse.com

Similarity Analysis of Court Judgements Using Association Rule Mining on Case Citation Data- A Case Study

Akhil M Nair

PG Scholar, Department of computer Science,
Christ University, Bangalore-560029, India.

Rupali Sunil Wagh

Associate Professor, Department of computer Science,
Christ University, Bangalore -560029, India.

Abstract

Information Retrieval System (IRS) is an automated mechanism of retrieving required information from a collection of unstructured or semi-structured data. IRS reduces the efforts of identifying the required information from an enormous database. Legal domain is one of the major producers of complex information which consist of semi-structured and unstructured data. Knowledge based legal information systems are revolutionizing all processes involved in this domain and hence need for more effective legal knowledge management approaches are increasing. This paper proposes association rule mining as knowledge extraction technique that can be used effectively for analyzing relatedness of documents in legal domain. Through this work, authors present their efforts in analyzing similarity in legal documents from the citations done in court judgement by applying Association rule mining.

Keywords: Knowledge management, legal domain, case similarity analysis, association rule mining, citation analysis.

I. INTRODUCTION

Legal domain is one of the major areas where a large amount of information is produced over time which leads to information overload. Legal domain contains the legal documents. Legal documents are collection of judgements given by the judges in different courts. In legal domain information overload has adverse effects in finding the

similar legal documents. The legal databases have numerous judgements from different domains. To find a similar case, stakeholder needs to browse through all documents in a database to find a set of similar judgements which is hence a time consuming and cumbersome task. With the prior experience and knowledge, a lawyer browses database and once he identifies an old judgement (say old_j) which satisfies the requirements, he searches for more judgements that are similar to old_j for analysis of those judgements. Since the number of judgements is enormous and judgements are huge and are complex, an automated mechanism to find the similar judgements based on their citations is a non-trivial problem.

Legal informatics is a field for effective management and organization of these legal documents. The legal domain contains both structured and unstructured data. Unstructured data in legal domain are the plain texts and facts in legal documents. Structured data can be referred to the citations in legal documents, dates, courts and case numbers. The judgements in common law system gives link to the judgements taken in similar previous cases. These links are also known as the citations. It gives an idea of the articles and judgements that were referred to take a judgement for a case. It also gives the idea of cases that fall under similar category. Citation analysis is used in legal domain to build case-citation network. Citations can be used to analyze hidden patterns in the judgements. Citations can also be used to analyze the relationship between the cited and the citing judgements.

In this paper an effort is made to analyze the similarity of the documents based on their citations. The study is done on legal documents for cyber related crimes. The crimes that are committed via internet are called the cyber-crimes. These crimes are committed against an individual or group of individuals with a criminal motive using the modern telecommunication systems. These cyber-crimes include phishing, email fraud, spams, hacking, illegal downloading etc. Indian parliament passed an act called the 'information technology Act 2000' on May 2000, to ensure the control over cyber-crimes. The dataset for the analysis belongs to the 'Information Technology Act 2000'.

Association rule can be defined as if/then statements, used to find hidden relationships between seemingly unrelated data. Application of association rule on citations of the judgements gives citations that frequently occur together. If two or more judgements have a set of citations that were cited frequently together in the dataset, then it can be inferred that they belong to a similar category and also that there exist relationship among cited frequent documents. Extraction of such relationship among legal documents may help in saving time and man labor while searching a database or library to find the similar cases and relevant legal documents.

1.1 LITERATURE REVIEW

Legal analytics is the process of acquiring valuable information from the pre-recorded data to assist legal practitioners. It involves extracting data present in legal documents, aggregating the data to give the judgements that were unknown previously. According to legal professionals, analytics though is a late entrant in legal domain is radically

changing legal education, legal proceedings in the courts and legal research [1]. Legal analytics uses advanced technologies like NLP, machine learning, to tidy-up, organize and analyze data from the legal documents. Legal analytics gives judges and lawyers an idea of how the judgements were taken for similar cases and also precisely how will the decision be taken in future insights [2][3].

Artificial Intelligence automates the process of problem solving and legal reasoning. Legal texts, documents and norms have their own logical characters which can be represented as facts based on their application of rules [4][5]. In the field of legal domain, intelligent systems do the data acquisition, building of knowledge models and the implementation of systems. There are different tools that are suited for the evaluation of the intelligent systems in legal domain. The classification system categorizes legal cases into most appropriate categories to build a knowledge model if needed. There is always an involvement of knowledge engineers who does the analysis to build the knowledge base. Intelligent systems automatically analyses and build the knowledge bases and hence minimizing the role of the knowledge engineers. These systems help the domain experts to build, organize and maintain the knowledge bases easily [6][7]. Since legal domain produces a huge amount of information, more intelligent techniques are to be used to provide a better grouping of the legal documents. Unsupervised text mining techniques have been used in this domain to facilitate the development of logic in legal documents by providing better and efficient insights into the available data [8][9].

Unstructured information holds the hidden relations that contain the links between different documents known as citations. Knowledge based legal information systems is need of the day. Citation analysis is a method to discover the hidden relationships between the documents and used to analyze the knowledge transfer from authors, articles and documents from various domains. Hence the study on citations of the legal documents becomes significant .Citation from a patent to a scholarly article provides information of the industrial and commercial application of that publication. The analysis of patent citation gives an idea of the relationship between science and industry. Since Google patents do not provide indexing for the academic citations, the process of bibliometric study on patent citations is time-consuming. It also does not support API searches and it becomes difficult to search in a large scale using its inbuilt interface. So, an automatic Bing search using the Bing crawl resolves the problem. It is in combination with the automatic filtering of results for the duplicate data produced by the Bing search. [10]

Similarity search and precedent search is widely used operation by legal professionals. There are many methods like legal-term cosine similarity and all-term cosine similarity that gives a better similarity result. The legal-term cosine method proves to be a better way to find the similarity. Vector space model is used in this method for similarity computation. For a legal cosine similarity, only terms that are there in the legal dictionary are considered to be the representative terms and for each term the tf-idf values are calculated. In all-term cosine method, instead of the terms in the legal dictionary, all the terms are considered to be the representative terms [11]. The Bibliographic similarity method is used with the number of the common out-citations.

Legal-term cosine and bibliographic similarity method proved to be the efficient methods to find the similar legal judgements. Network analysis in legal domain has the ability to bring about major change in the analysis and management of legal information [12]. It is an emerging research field which aims at representing legal concepts as nodes and relations among them as edges. This approach has been very popular among research community both in information technology as well as law. Several scholars noticed that the law citation contain valuable information that gives us the relevance of the previous judgements. [13] Citations can be analyzed as citation networks. Case citation networks are scale-free that means very small number of cases receive most citations. The network analysis done on the Canadian Law shows that the in-degree centrality and PageRank scores of case law within the Canadian Law database are effective predictors of how frequently those cases will be viewed on the websites. Plotting the network ranking of a case overtime and determining the slope can help to pinpoint the most influential cases [13]. Citations can also be analyzed independently as structured data for more insights into the relatedness of legal documents and legal issues. This relation can be explored using association rule mining. The advantage of association rule lies in its simplicity in providing outcomes in the form of readily understandable rules which can then be analyzed, evaluated and examined. Basic association rule mining algorithms produce all possible patterns from a given dataset. The produced pattern is then pruned to filter out the non-frequent patterns and generate rules for the frequently occurring patterns. These methods are time consuming and require more computational resources. In order to minimize the time and computational requirements, finding the maximum co-occurring pattern without finding all the possible frequent pattern combinations at once and filtering them out later will be a more efficient method. This approach will have significantly less time consumption and less usage of computational resources [14]. Association rule is utilized to distinguish the relation amongst hypertension and different modifiable hazard factors, for example, smoking or drinking, trying to build up a hypertension-management program [15].

II. METHODOLOGY

Methodology used for the study has three major steps - Data collection, Data Pre-processing and Building and evaluating the knowledge model. This can be represented using figure 1:

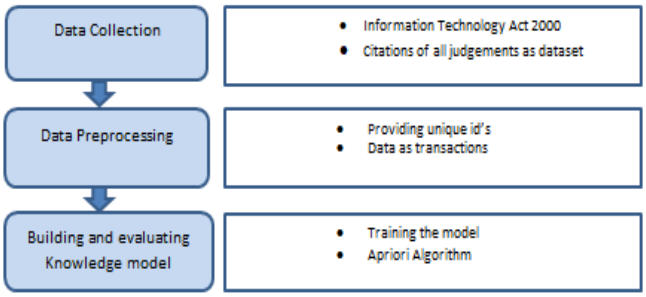

Fig.1 Methodology

2.1 DATA COLLECTION

Citation dataset used in this paper is made from the online database 'indiankanoon.org'. Citations of all the cases under the Information Technology Act 2000 conducted in different high courts in different states of India are taken from this website. The dataset was first made in a case and citation format. The column 'cases' had all the cases under information technology act 2000 and the second column contained all the citations cited by the corresponding cases. The dataset had 597 observations.

2.2 DATA PREPROCESSING

The preprocessing and analysis of the dataset are done using R Studio. The first step of preprocessing removed all the unwanted terms from the citations and cases for appropriate output. Secondly the unique id's for the cases and citations are given to get more detailed information. If there is a citation that is being repeated the id for the citation is used. Then the two column format is converted into a row-column format where the rows are the cases and number of columns represents the number of citations with the citation ids'. The dataset is then transformed into transactions (into a sparse matrix) for applying the algorithm. The purpose of converting the original dataset into transaction data is to remove broken data across multiple rows since the number of citations may differ in each case.

2.3 BUILDING AND EVALUATING KNOWLEDGE MODEL

The transaction dataset is trained using the Apriori algorithm. Association rule gives an idea of how frequently a citation occurs in combination with another citation. The support measure in association rule gives idea of how frequently a citation appears in the database and confidence is the measure of number of times the if/then statements have been found true. Another important value in association rule is lift. It is the confidence of a rule divided by the expected confidence.

Once the dataset is in transaction format apriori algorithm is applied on the data specifying the support and confidence value. This algorithm is used for finding pattern of frequent itemsets in the dataset and the association rule that defines the pattern.

- Apriori algorithm mines frequent itemsets for Boolean rules.
- This algorithm uses a bottom-up approach where frequent subset is extended one at a time.
- It is designed to apply on transaction data example. Information of website frequentation.
- The property of Apriori says that – any subset of frequent itemsets is also frequent.

The Apriori algorithm performs a level-wise search. For the first iteration it counts the entire 1-itemsets find the count that is more than the threshold. Then it combines those itemsets to form 2-itemsets, counts them to and checks for large 2-itemsets [16].

III. RESULTS AND DISCUSSIONS

Fig 2 shows the list of 20 rules with their support and confidence value with the proper count of how frequently the itemset occurred in the dataset.

- From the list it is clear that most frequently occurred itemset is {1104} →{1117} in the dataset.

	lhs		rhs	support	confidence	lift	count
[1]	{1104,1106}	=>	{1103}	0.04288165	0.9615385	19.330239	25
[2]	{1103,1104}	=>	{1106}	0.04288165	0.8928571	18.590561	25
[3]	{1106}	=>	{1103}	0.04288165	0.8928571	17.949507	25
[4]	{1103}	=>	{1106}	0.04288165	0.8620690	17.949507	25
[5]	{1104,1117}	=>	{1103}	0.04459691	0.8666667	17.422989	26
[6]	{1103,1106}	=>	{1104}	0.04288165	1.0000000	16.657143	25
[7]	{1106,1117}	=>	{1104}	0.04116638	1.0000000	16.657143	24
[8]	{1104,1117}	=>	{1106}	0.04116638	0.8000000	16.657143	24
[9]	{1103,1117}	=>	{1104}	0.04459691	1.0000000	16.657143	26
[10]	{1103}	=>	{1104}	0.04802744	0.9655172	16.082759	28
[11]	{1104}	=>	{1103}	0.04802744	0.8000000	16.082759	28
[12]	{1106}	=>	{1104}	0.04459691	0.9285714	15.467347	26
[13]	{1029,881}	=>	{882}	0.04116638	0.8275862	14.190669	24
[14]	{1029,882}	=>	{881}	0.04116638	0.8000000	11.375610	24
[15]	{1103,1104}	=>	{1117}	0.04459691	0.9285714	10.827143	26
[16]	{1104,1106}	=>	{1117}	0.04116638	0.9230769	10.763077	24
[17]	{1103}	=>	{1117}	0.04459691	0.8965517	10.453793	26
[18]	{1106}	=>	{1117}	0.04116638	0.8571429	9.994286	24
[19]	{1104}	=>	{1117}	0.05145798	0.8571429	9.994286	30
[20]	{881,882}	=>	{1029}	0.04116638	0.8888889	6.909630	24

Fig. 2 List showing 20 rules

- The rules for the support value 0.04 and confidence value 0.80 is shown in Fig. 3. Each rule is represented with a circle in the figure. In the graph, confidence level is represented with the size of the circle and lift value is represented with the intensity of the color of the circle.
- The rules generated for 12 of the itemsets is shown in Fig. 4. The lhs and rhs simply mean left hand side and right hand side. Table clearly tells us that 4.1 % (support) transactions comprises {881,882} in the dataset. The confidence value tells that for combination {881,882} there is 88.8 % possibility that the citation 1029 is also cited with it.
- Citation 881 is about fraudulent usage a forged document or electronic record as a genuine one. Citation 882 is about forging a document or electronic record purposefully for cheating purpose. The citation 1029 is about cheating and delivering a property dishonestly . Hence the citations have similarity since they tell about fraudulent and purposeful cheating.
- The case ids' 428 (*Veerdavinder Singh And Anr vs State Of Punjab Etc on 24 February, 2015*) and 440 *(Harjot Kaur vs State Of Punjab on 8 December, 2015*) cites {881, 882, 1029}, therefore, based on the rule, we can say that the cases 428 and 440 are similar and can be considered to fall under cheating and fraudulent activities.

- Similarly from the figure 3, rule says that the citation {1117} is associated with the citation {1103, 1104}. From Fig. 3 it is clear that for the citation {1103, 1104} there is 92.85% possibility that {1107} is also cited along with it.
- The citations 1103, 1104 and 1107 are the sections 468,467 and 471 of IPC respectively that tells about fraudulent activities. Therefore the cases that cite these citations may belong to a similar category or have some elements that the cases may have in common.

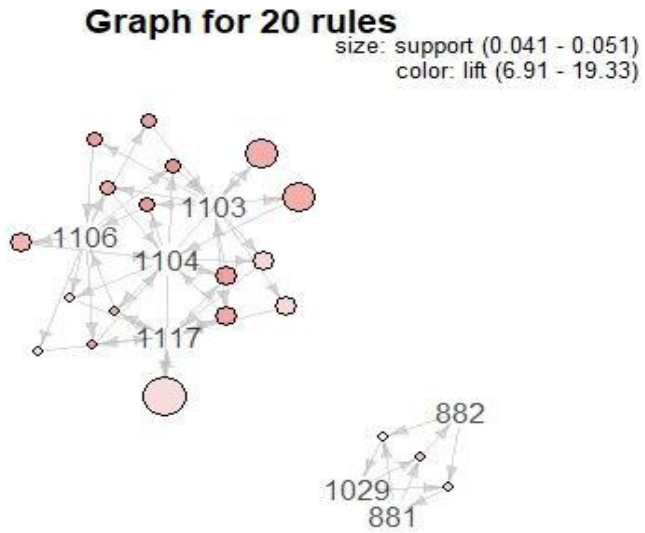

Fig. 3 Graph showing 20 rules

	lhs	rhs	support	confidence	lift
16	{1103,1104}	{1117}	0.04459691	0.9285714	10.827143
13	{1104,1106}	{1117}	0.04116638	0.9230769	10.763077
7	{1103}	{1117}	0.04459691	0.8965517	10.453793
1	{1106}	{1103}	0.04288165	0.8928571	17.949507
12	{1103,1104}	{1106}	0.04288165	0.8928571	18.590561
19	{881,882}	{1029}	0.04116638	0.8888889	6.909630
9	{882}	{1029}	0.05145798	0.8823529	6.858824
18	{1104,1117}	{1103}	0.04459691	0.8666667	17.422989
2	{1103}	{1106}	0.04288165	0.8620690	17.949507
4	{1106}	{1117}	0.04116638	0.8571429	9.994286
8	{1104}	{1117}	0.05145798	0.8571429	9.994286
21	{1029,881}	{882}	0.04116638	0.8275862	14.190669

Fig. 4 Table showing rules of {881,882} -> {1029}

IV. CONCLUSION

In this paper association rule mining is used to find the similarity between cases using the citations. Apriori algorithm generates a set of rules that gives the idea of how frequently a case is being cited in other cases in the dataset. From the result we can deduce that two or more judgements with citations that frequently occur together are similar. Hence these cases can be assumed to be of the same category. It can also be inferred that the citations occurring together frequently are also related with each other. Citation based association Rule Mining approach can be used in similarity searches in legal databases and digital legal libraries to get the relevant documents.

The future work lies in finding rules using other association rule mining algorithms like the AIS, FP growth, Apriori Hybrid for higher efficiency, accuracy and also to improve the performance of the system. This work focuses on one subdomain of law. It can further be extended to cover more domains to understand interrelationships of legal issues across subdomains of law.

REFERENCES

[1] K. Lee, S. Azyndar, and I. Mattson, A New Era: Integrating Todays Next Gen Research Tools Ravel and Casetext in the Law School Classroom, *SSRN Electronic Journal*, 2015.

[2] O. Byrd, Legal Analytics vs. Legal Research: What's the Difference? *Law Technology Today, 12-Jun-2017. [Online]*. Available: http://www.lawtechnologytoday.org/2017/06/legal-analytics-vs-legal-research/.[Accessed: 15-Nov-2017]

[3] "Legal Analytics," *Argopoint*. [Online]. Available: http://www.argopoint.com/legal-analytics. [Accessed: 15-Nov-2017].

[4] L. K. Branting, "Data-centric and logic-based models for automated legal problem solving," *Artificial Intelligence and Law*, vol. 25, no. 1, pp. 5–27, 2017.

[5] F. Bex, H. Prakken, T. V. Engers, and B. Verheij, Introduction to the special issue on Artificial Intelligence for Justice (AI4J), *Artificial Intelligence and Law*, vol. 25, no. 1, pp. 1–3, 2017.

[6] A. Stranieri and J. Zeleznikow, Tools for intelligent decision support system development in the legal domain, *Proceedings 12th IEEE Internationals Conference on Tools with Artificial Intelligence. ICTAI 2000*.

[7] E. Bellucci and J. Zeleznikow, AI techniques for modelling legal negotiation, *Proceedings of the seventh international conference on Artificial intelligence and law - ICAIL 99*, 1999.

[8] Rupali Sunil Wagh, Exploratory analysis of legal documents using unsupervised Text Mining techniques, *International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 2, February – 2014*

[9] Rupali Sunil Wagh, Knowledge discovery from legal documents dataset using text mining techniques, *International Journal of Computer Applications (0975 – 8887),March 2013.*

[10] Kousha, K. and Thelwall, M. (2017), Patent citation analysis with Google. *J Assn Inf Sci Tec, 68: 48–61. doi:10.1002/asi.23608.*

[11] S. Kumar, P. K. Reddy, V. B. Reddy, and A. Singh, Similarity analysis of legal judgments, *Proceedings of the Fourth Annual ACM Bangalore Conference on - COMPUTE 11*, 2011.

[12] Aaron Abood, Dave Feltenberger, Automated Patent Landscaping, *Proceedings of the Workshop on Legal Text, Document, and Corpus Analytics, LTDCA 2016.*

[13] Neale, Thom. Citation Analysis of Canadian Case Law, *Journal of Open Access to Law* vol. 1, no. 1 (2013.

[14] Ömer M. Soysal, Association rule mining with mostly associated sequential patterns, *Expert Systems with Applications*, vol. 42, no. 5, pp. 2582–2592, 2015.

[15] Y. M. Chae, S. H. Ho, K. W. Cho, D. H. Lee, and S. H. Ji, Data mining approach to policy analysis in a health insurance domain, *International Journal of Medical Informatics*, vol. 62, no. 2-3, pp. 103–111, 2001.

[16] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, Dynamic itemset counting and implication rules for market basket data, *Proceedings of the 1997*