Using Regular Expressions, Word Embedding and Topic Modeling to Find Cited Legal Documents

Yogesh H. Kulkarni
Consultant, RightSteps Consultancy,
Pune, India.
yogeshkulkarni@yahoo.com

Rishabh Patil
Engineer, RightSteps Consultancy,
Pune, India.
rishabh@rightstepsconsultancy.com

Srinivasan Shridharan
Founder, RightSteps Consultancy,
Pune, India.
srini@rightstepsconsultancy.com

ABSTRACT

Indian judicial system follows “Common Law System” which refers to written-law(statute) as well as past cases (precedent) to give verdict on the legal cases. Due to thousands of past cases it becomes tedious and error-prone to find relevant precedent, manually. An automatic precedent retrieval system is the need of the hour. One of the key requirements of such information system is to find cases which could be “similar” to the case in hand. The “similarity” used in this paper is about citations. The problem is of predicting prior cases which could potentially be cited by a particular case text. This paper proposes such association system using mixed approaches. It employs rule-based Regular Expressions based on references to statute and Articles. It finds cosine similarity between cases using vectors generated by popular word embedding called doc2vec. It also leverages topic modeling by finding matches between cases based on the number of common topic words. Towards end, this paper demonstrates efficacy of the proposed system by generating cite-able documents from test data-set.

CCS CONCEPTS

• **Information systems → Document topic models**; • **Computing methodologies → Information extraction**;

KEYWORDS

Information Retrieval, Regular Expressions, Word Embedding, Topic Modeling, Legal, Word2Vec, Citations, NLP.

ACM Reference Format:
Yogesh H. Kulkarni, Rishabh Patil, and Srinivasan Shridharan. 2017. Using Regular Expressions, Word Embedding and Topic Modeling to Find Cited Legal Documents. In *Proceedings of Forum for Information Retrieval Evaluation (Fire-IRLeD’17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

In “Common Law System”, both, statutes(the “written law”) and precedent (the “prior cases”) are deliberated upon equally while deciding on the judgments. This helps being uniformity of judgments across similar circumstances. Due to vast number of prior cases, which are increasing at rapid pace, it is impossible to employ manual methods to retrieve relevant precedents. Automatic Precedent Retrieval System (APRS) is desirable in such situation. One of the key functionality necessary in APRS is to find “similar” cases so that they can be cited or referred from the case being built. The notion of “similarity” has various connotations. In the context of the given problem, it is said to be the documents which share citations. In other words, the task is to find such prior cases which are potentially cite-able from the case in hand.

Legal texts are typically lengthy and unstructured in nature. It is challenging to find similarity score among two texts by just counting words or their frequency distributions or such preliminary statistical measures. Need to embed higher level constructs such as word embedding to introduce semantic similarity as well as higher level clusters given by Topic Modeling.

The aim of this paper is to propose automatic cite-able texts detection-prediction system for legal texts. It is unsupervised (no labeled training data) technique comprising of Regular Expressions, Word2Vec and Topic Modeling. All being employed to give similarity score based on different aspects of the texts. Final rank is arrived at using weighted sum of the individual scores.

The contributions made in this paper are as follows:

(1) Detecting statute and Articles based on Regular Expressions.
(2) Proposing cosine similarity between texts based on vector generated by Word Embedding (word2vec/doc2vec).
(3) Proposing Document-Topics-Words distribution for all texts and then scoring similarity based on common topic-words.

The paper has been structured as follows: in the following section 2 the catchphrases detection task has been described in details, as definition of the problem. In section 3, structure of the training data has been explained. Next, the proposed system is elaborated in section 4. It describes ranking method based on weighted-sum of scores from individual methods. Section 5 discusses the findings drawn from this work.

2 TASK DEFINITION

Legal cases typically cite statute, Articles and previous cases relevant to them. Thus, it is necessary to form association or similarity between documents based on citations, so that it can be leveraged for Precedent retrieval. Given sets of current and prior cases, the task is: *For each document in the first set, the participants are to form a list of documents from the second set in a way that the cited prior cases are ranked higher than the other (not cited) documents.*[2]

3 DATA-SET

Fire-2017 [2] data-set contains following directories:

Fire-IRLeD’17, Dec 2017, IISc, Bangalore, India

ACM ISBN 123-4567-24-567/08/06…$XX.00
https://doi.org/10.475/123_4

(1) Current cases: A set of cases for which the prior cases have to be retrieved, $current\_case_{<}i > .txt$ where $i = 0001 \rightarrow 0200$. Sample document looks like "**Judgment** IN THE SUPREME COURT OF INDIA....".
(2) Prior cases: contains prior cases that were actually cited in the case decision along with other (not cited) documents, $prior\_case\_ < i > .txt$ where $i = 0001 \rightarrow 2000$. Sample document looks like " 551; AIR 1996 SC 463; 1995 (6) SCC 315; 1995 (7) JT 225; 1995 (5) SCALE 690 (11 October 1995) JEEVAN REDDY, B.P. (J) JEEVAN REDDY, ...the assessee. No costs".

4 SYSTEM DESCRIPTION

The proposed APRS solution uses the following 3 distinct approaches to determine similarity. The final similarity score is computed as a weighed average of the scores generated by these 3 approaches.

4.1 Regular Expressions based

Cases refer to statue in the form of legal Articles, such as Article 270, 370, etc. Such references can be extracted using Regular Expressions. In this paper, patterns like `r' article (\d+)'` are used for both, current as well as prior cases. For a current case, all the prior cases are collected which have same Articles.

4.2 Topic Modeling based

Basic premise of this approach is that if most of the topics extracted from documents match then they are similar or cite-able. In this paper, Document-Topics-Words distribution is generated by Latent Dirichlet Allocation (LDA) algorithm in gensim library[3]. Score of similarity is calculated based on ratio of matching topic-words to the total.

4.3 Doc2Vec Similarity based

Word2vec has emerged one of the most popular vectorization based on semantic similarity [1]. Process to generate document vectors (based on word2vec) was:

(1) Got every case as cleaned text, split it to form list of words/-tokens, for both, current and prior cases.
(2) Created gensim TaggedDocument for each case text, giving filename as tag.
(3) A Map of tag to the content i.e. word-list for each cases were generated and saved for reuse.
(4) LDA model was built and saved. It was used to generate document vectors for both current and prior cases.

A similarity matrix was generated where current cases are rows and prior cases as columns with values as cosine similarity between document vectors of the current-prior case pair (row-column). The values act as score for this particular approach.

4.4 Results

Each current-prior case pair has a final score based on weighted sum of scores from individual approaches mentioned above. Due to lack of labeled training data, the weights were decided heuristically. The results were presented as sorted list of prior case for each current case, and looked as follows:

Current case	Prior case	Rank	Score
current_case_0001	prior_case_0780	0	1.783
current_case_0001	prior_case_1256	1	1.743
current_case_0001	prior_case_0838	2	1.727
...			
current_case_0116	prior_case_1533	0	1.003
current_case_0116	prior_case_0411	1	0.929
current_case_0116	prior_case_1600	2	0.877
...			

Table 1: Submission file

Results on the test cases were evaluated by Fire-2017[2] and conveyed as below:

Mean Average Precision	Mean Reciprocal Rank
0.2017	0.45

Table 2: Results conveyed by FIRE[2]

5 CONCLUSIONS

In this paper, a brief overview of Automatic Citation Prediction System was presented to discover cite-able prior cases. It was found that the problem of citation detection needs to be modeled as a mixed approach, employing rule based, machine learning and deep learning based approaches rather than a simple cosine similarity of tf-idf (term frequency inverse document frequency) approach. Weighted sum of scores by individual approaches was done. A threshold cut-off was decided to prune out irrelevant cite-able prior cases. Current cases and their predicted cite-able prior cases were presented along with corresponding ranking scores.

As a future work, if sufficient gold standard data is available, one can explore more into word embedding approach so as to find cite-able cases based on pure semantic similarity.

VITAE

Yogesh H. Kulkarni works as Data Science Consultant and Trainer. Profile: https://www.linkedin.com/in/yogeshkulkarni/
Rishabh Patil works as Data Engineer. His profile is at https://www.linkedin.com/in/rishabh-patil-256a25124/.
Srinivasan Shridharan is Data Scientist and entrepreneur. Profile: https://www.linkedin.com/in/srinivasan-shridharan-08a86a6/.

ACKNOWLEDGMENTS

Wish to thank Ankur Parikh, a keen researcher of Deep Learning and NLP, for discussions.

REFERENCES

[1] Chris McCormick. 2016. Word2Vec Tutorial - The Skip-Gram Model. (Apr 2016). http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
[2] Information Retrival Society of India. 2017. Forum for Information Retrieval Evaluation. (Dec 2017). https://sites.google.com/view/fire2017irled/track-description
[3] Radim Rehurek. 2017. Gensim: Topic Modeling for humans. (Sep 2017). https://radimrehurek.com/gensim/