文章编号：2095-2163(2021)06-0001-05 中图分类号：TP399 文献标志码：A

基于法律要素引导的相似案例推荐算法

刘博阳，李 尚，叶 麟，张宏莉
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：大量的法律案例以数字化的形式存储，使得法务工作者及普通民众可以轻松地从法律案例库中搜索需要的信息，其中有用但又很有挑战的一项任务就是相似案例推荐。为了准确地从法理角度推荐相似案例，本文提出了一个基于神经网络的相似案例推荐模型，该模型首先用法律要素引导每个案例的文本表示向量的生成，进而用生成的向量计算任意一对案例的相似度分数，将相似度最高的案例集合作为推荐的相似案例。在真实的数据集上的实验证明本文的模型优于常用的文本相似度计算模型。

关键词：文本表示；文本相似度；法律案例；神经网络

Similar case recommendation algorithm based on legal elements

LIU Boyang, LI Shang, YE Lin, ZHANG Hongli
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 A large number of legal cases are stored in digital form, which allows legal workers and ordinary people to easily search the required information from the legal case database. One of the useful but challenging tasks is the recommendation of similar cases. In order to accurately recommend similar cases from a legal perspective, this paper proposed a neural network based similar case recommendation model. The model first uses legal elements to guide the generation of the text representation vector of each case, and then uses the generated vector to calculate the document similarity score of any pair of cases. The set of cases with the highest similarity is regarded as the recommended similar cases. Experiments on real data sets prove that our model is better than some commonly used models.

【Key words】 legal case; document similarity; text representation; neural networks

0 引 言

随着法律知识的普及，人们习惯于使用法律手段解决问题，导致了需要被解决的各类案件数目逐年增长，这无疑给法律工作者带来了巨大的压力。近些年，有很多学者开始研究法律领域的人工智能，在判决预测，案件分类等方面取得了很多成就。然而相似案件推荐的算法研究较少，由于相似的案件往往有着相似的判决结果，寻找相似案例对法务工作者乃至普通大众都有参考意义，因此本文着眼于此，提出了一种基于法律要素的模型来提高相似案例推荐的准确性。

相似案例推荐，即在给定一个判决文书的情况下，找出在法律角度上与之相似的案件。从本质上来说，此问题还属于文本相似度的研究，然而案件的相似并不是简单的文本相似，而是挖掘文本中所包含的法律要素的相似，这就导致了单纯使用文本相似度计算方法并不能找到高度相似的法律案件。

综上，本文提出了一种由要素引导的神经网络模型来形成案件文本的向量表示，然后利用该向量表示计算两两案件的余弦相似度，并返回相似度最高的一系列案例作为给定案件的相似案例。本文在真实的故意伤害罪数据集上进行了实验，发现返回的相似案例要优于常用的文本相似度计算方法。

1 相关工作

在欧美等实施判例法的国家中，每一个案件的判决中都会引用以往的判决案件作为新案件判决的依据，因此，案件之间就会构成一个引文网络。Opijnen 使用网络度统计和结构属性提取相关文件的法律域名领域[1]；Wagh 等人提出了利用案例引证网络节点的中心性和介数性来寻找印度法院判决

基金项目：国家重点研发计划(2018YFC0830602)。

作者简介：刘博阳(1994-)，男，硕士研究生，主要研究方向：自然语言处理、文本挖掘、机器学习等；李 尚(1989-)，男，博士研究生，主要研究方向：机器学习、自然语言处理、信息安全等；叶 麟(1982-)，男，博士，副教授，主要研究方向：网络安全、P2P 网络、网络测量和云计算等；张宏莉(1973-)，女，博士，教授，博士生导师，主要研究方向：网络与信息安全、网络测量与建模、网络计算等。

收稿日期：2020-10-05

相似性的方法[2]；Minocha 等人定义了一个法律角度的离散度，用于衡量两个案例的相邻案例集合的相似度，进而发现引文网络中相似的案例[3]。

基于引文的相似性的法院案件在法律领域无疑具有非常重要的意义，但是案例引证图通常非常稀疏，因此基于机器学习和自然语言处理的方法被提出[4-5]；Ashley 等人利用基于案件特征的最近邻算法计算案件相似度[6]；Carneiro 等人在采用基于词频的贝叶斯统计方法对法律案例的相似度进行计算[7]。

随着 word embedding 的出现，信息检索已经转向了神经信息检索。Xia 等人利用法律的文本语料库训练 word2vec 模型，用于计算法律文本的相似度[8]；Vo 也表示基于词嵌入的文本语义表示对法律文本检索领域很有帮助[9]。但这些方法都忽视了对法律领域知识的使用，因此本文从法律要素的角度将领域知识结合到文本的向量表示中，使得模型能够从法律要素的层面寻找相似案例。

2 模型构建

2.1 模型框架

本文提出了相似案例推荐模型，该模型的整体框架如图 1 所示。首先将案例输入到一个神经网络中，用案件中包含的法律要素作为标签训练网络，并利用输出层前一层输出的向量作为案例的向量表示，由于在训练过程中包含了语义信息和要素信息，因此利用该向量可以很好的表示一个案件的语义信息和案件中所包含的法律要素情况，本文利用该向量计算任意两个案件之间的余弦相似度，最终返回相似度最高的 K 个案件最为推荐的相似案例。

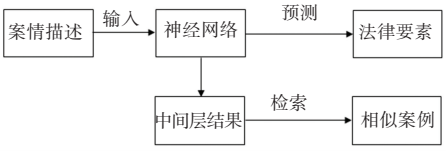

图 1 相似案例推荐模型整体框架图

Fig. 1 Overall framework of similar case recommendation model

2.2 法律要素预测模型

预测要素标签的神经网络结构，如图 2 所示。整个网络分为 5 层，分别为词嵌入层、语义嵌入层、两个全连接层和一个输出层。

词嵌入层的输入是经过分词，去除停用词等预处理操作之后的文本，该文本可以表示为式(1)：

$$d = \{w_1, w_2, \cdots, w_i, \cdots, w_n\}, \quad (1)$$

其中，w_i 代表分词后的短语或词组。经过词嵌入层后，w_i 会被映射到语义空间中形成向量，式(2)：

$$D = f_{emb}(d), \quad (2)$$

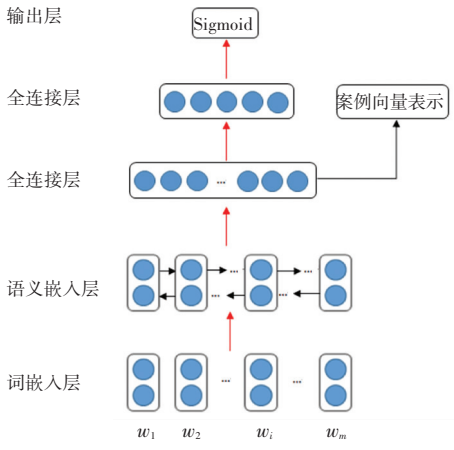

图 2 法律要素预测模型

Fig. 2 Legal element prediction model

在这一步中，词向量的转化采用 word2vec 实现。进一步地，为了使文本能够学习上下文的语义信息，本文采用了基于循环神经网络的语义嵌入层，式(3)：

$$h = f_{rnn}(D), \quad (3)$$

在接下来的两个全连接层中，第一个全连接层的输出维度即案件的向量表示维度，该层的输出不仅被当作下一个全连接层的输入，也用于形成一个案件的向量表示，被进一步用于相似案例推荐，式(4)：

$$fc = \sigma(W_1 D + b_1), \quad (4)$$

第二个全连接层的输出维度与法律要素的类别数相同，用于预测法律要素，式(5)：

$$y = \sigma(W_2 fc + b_2), \quad (5)$$

由于预测法律要素属于多标签分类问题，本文在输出层中采用 *sigmoid* 函数，式(6)：

$$p = sigmoid(y), \quad (6)$$

本文使用的损失函数为交叉熵损失函数，其计算公式(7)为：

$$Loss = -\sum_k y_k \log p_k + (1 - y_k) \log (1 - p_k). \quad (7)$$

2.3 法律要素选择

本文根据故意伤害罪相关法律条文中规定的法律要素和数据集中法律要素出现的次数，选取了出现次数较多的前 35 个要素作为标签，这些标签都是结合人工观察与正则表达式抽取出来的，具体的法律要素见表 1。

表 1　故意伤害罪相关法律要素

Tab. 1　Legal elements of intentional injury

被告人相关要素	孕妇、未成年人、残疾人、老年人、共同犯罪、精神病人、初犯、偶犯、前科、累犯
故意伤害罪相关要素	互殴、故意伤害他人身体、击打被害人头部、持刀具等利器伤人、持棍棒等钝器伤人、暴力索财致人伤害、因婚姻家庭纠纷引起的故意伤害、因锁事纠纷、轻伤一人、轻伤多人、重伤一人、重伤多人、轻微伤、轻伤二级、轻伤一级、重伤二级、发生口角
补偿行为相关要素	积极主动赔偿、积极抢救被害人、坦白、悔罪、立功、自愿认罪、自首、取得被害人谅解

3　实　验

3.1　数据集

本文中使用的数据集是从中国裁判文书网中爬取的，该数据集中一共有 2 148 篇判决文书，其罪名均为故意伤害罪。在寻找相似案例的时候，本文使用的并不是完整的判决文书作为输入，而是使用判决文书中案情描述的部分作为输入。

3.2　基线方法

在预测法律要素的模型中，本文比较了 RNN 模型及其几种常见的变体，包括 LSTM、GRU、BiGRU 以及采用 attention 机制的 BiGRU。在神经网络的训练中，隐藏层的输出维度均为 128，均采用 adam 作为优化器，学习率设置为 0.001，第一个全连接层的输出为 200 维。

在相似度计算的对比模型选取中，本文主要采用了 4 种常用的用于计算文本相似度的无监督模型，tf－idf、word2vec、doc2vec 和结合 tf－idf 的 word2vec 作为对比实验，其中结合 tf－idf 的 word2vec 是将一个单词的 tf-idf 的值作为 word2vec 的权重，进而获得句子的向量表示。

3.3　评价指标

由于缺乏刑事类相似案件的数据集，并且司法领域也并没有明确的规定表明满足什么条件的两个案件可以称为相似案件，即相似案例的判定并不存在一个充要条件。但由于中国司法领域对类案类判的要求，相似案例的判决结果应当相似，即判决结果的相似是相似案例的必要条件。因此，本文拟采用这个必要条件对模型进行测评。在司法领域中，判决结果主要是罪名、相关法律条文和刑期，本文分别给出这 3 个方面的相似度计算公式。

对于罪名而言，相似案例都具有相同的罪名，罪名维度的相似性计算公式可以表示为式(8)：

$$sim\_c(C_1, C_2) = \begin{cases} 1, & \text{if } C_1 = C_2, \\ 0, & \text{else.} \end{cases} \tag{8}$$

由于法律条文是一个集合，这里借用杰卡德系数来计算两个案件法律条文维度的相似性，式(9)：

$$sim\_a(A_1, A_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}. \tag{9}$$

刑期都是采用整数表示的，为了便于计算，本文将刑期的单位统一成月份，刑期相似度计算函数为式(10)：

$$sim\_t(T_1, T_2) = \frac{1}{\dfrac{|T_1 - T_2|}{\max(T_1, T_2)} + 1}, \tag{10}$$

最终，两个案件的相似度的综合计算为公式(11)：

$$sim(D_1, D_2) = sim_c * (sim_a + sim_t)/2, \tag{11}$$

由于本文使用的数据集都是在同一罪名下的数据，因此式(11)变为式(12)：

$$sim(D_1, D_2) = (sim_a + sim_t)/2, \tag{12}$$

相似案例推荐属于信息检索范畴，本文采用信息检索领域常用的 DCG(Discounted cumulative gain)作为评价指标，其基本思想是对信息检索返回的 p 个结果分别进行打分，并将这些结果的分数求和，得到 p 个返回结果的综合得分。该指标主要有两点假设，第一是返回的结果中，越相关的结果排在越前面越好，另一点是打分高的结果比打分低的结果好。该指标的计算公式(13)如下：

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log(i+1)}. \tag{13}$$

其中，p 代表返回的相似案例的个数；i 代表返回案件的顺序；rel_i 代表返回的第 i 个相似案例与给定案例的相似度程度；这里的 rel_i 即为本文中用于计算两个案件相似性得分的 $sim(D_1, D_2)$ 的值。

3.4　实验结果

首先给出法律要素预测模型的准确率、召回率和 $F1$ 分数，其具体数值见表 2。

表 2　法律要素预测模型效果

Tab. 2　Experimental results of legal element prediction

模型	准确率	召回率	$F1$ 分数
RNN	0.806 0	0.508 0	0.623 2
LSTM	0.721 7	0.561 9	0.631 8
GRU	0.731 4	0.565 1	0.637 6
BiGRU	0.739 9	0.578 4	0.649 3
BiGRU+attention	0.790 5	0.592 1	0.677 0

本文比较了 5 种不同的语义嵌入层模型，发现 BiGRU+attention 的效果最好，因此本文采用该模型

生成的向量作为案例文本的向量表示。

本文的模型与基线模型的效果对比见表 3。其中,DCG@5,DCG@10,DCG@20 分别代表返回的前 5 个,前 10 个和前 20 个案例的 DCG 的值。可以看出,本文模型在寻找相似案例的效果上比其它模型有明显的提高。

本文中模型对给定案例返回的前 5 个相似案例的结果见表 4。其中,粗体下划线的部分为案件涉及到的法律要素,返回的案例中包含的法律要素基本与给定案例相符合,其中返回的第二个案例与给定案例有着极高的相似性,所有返回案例涉及到的相关法律条文有一定误差,但刑期误差都较小。

表 3 相似案例推荐模型效果

Tab. 3 Experimental results of similar legal case recommendation

模型	DCG@5	DCG@10	DCG@20
Word2vec	1.077 2	1.605 1	1.932 6
Doc2vec	1.206 8	1.844 6	2.396 5
TF-IDF	1.498 2	2.241 2	3.189 5
TF-IDF+Word2vec	1.894 1	2.734 8	4.038 5
基于法律要素	2.266 5	3.317 4	4.856 5

表 4 某具体案例的相似案例推荐情况

Tab. 4 Similar legal case recommendation for a specific case

	案情描述	法律条文	刑期
给定案件	2014 年 10 月 4 日 19 时许,被告人刘某在本市丰台区右外翠林一里 1 号楼 1103 号内,因琐事与被害人兰某发生口角,后被告人刘某将被害人郎某打伤,致其左侧 9-11 肋骨共三处骨折,经法医鉴定被害人郎某的身体损伤程度为轻伤二级。本案民事赔偿部分已和解解决,被告人刘某一次性赔偿被害人郎某经济损失人民币五万元,被害人郎某对被告人刘某表示谅解。	[234,67,72,73,61]	6
1	2013 年 12 月 15 日 15 时 50 分许,被告人王某伙同“四哥”(在逃)在本市海淀区红五星歌厅内因琐事纠纷对被害人朱某进行殴打,致被害人朱某双侧鼻骨骨折,左侧额突骨折等伤,经鉴定为轻伤二级,2014 年 2 月 17 日,被告人王某被公安机关电话传唤到案,后如实供述了上述犯罪事实,经本院调解,被告人王某赔偿被害人朱某因伤造成的经济损失共计人民币 2 万元,被害人朱某对被告人王某的行为表示谅解。	[234,67]	8
2	2016 年 10 月 7 日 22 时许,被告人陆某在本市海淀区南平庄嘎子汽修厂门前路边,因琐事与被害人张某(男,41 岁)发生纠纷,并用拳击打被害人鼻面部,致其左侧鼻骨骨折,左侧上颌额突骨折等伤,经鉴定为轻伤二级。被告人陆某于 2016 年 11 月 16 日被公安机关电话传唤,后如实供述了上述犯罪事实。案发后,被告人陆某在其亲属的协助下赔偿被害人现金人民币十六万元,并获对方谅解。	[234,67,72,73]	6
3	2015 年 9 月 26 日 8 时许,被告人杨某在本市海淀区苏家坨镇温阳路加油站对面仓库因琐事与刘某(男,51 岁)发生口角并互殴。其间,被告人杨某用拳头击打被害人刘某面部,致其右眶内壁、上臂骨折等伤,经鉴定为轻伤一级。被告人杨某经电话传唤于 2015 年 9 月 30 日主动投案,后如实供述了上述犯罪事实。案发后,被告人杨某已赔偿被害人刘某人民币 3 万元,被害人刘某表示谅解。	[234,67,72,73]	8
4	2015 年 10 月 23 日,被告人杨某 1 在本市海淀区香山丰户营碧浪园歌厅内,因琐事与被害人杨某 2(男,52 岁)发生争执,后杨某 1 用牙齿将被害人杨某 2 右手环指咬伤,致被害人右手环指近节指骨骨折,经鉴定属轻伤二级。被告人杨某 1 于 2015 年 11 月 8 日向公安机关投案,后如实供述了上述犯罪事实。被告人杨某 1 赔偿被害人杨某 2 人民币 1.2 万元,取得被害人谅解。	[234,67,72,73]	6
5	2015 年 2 月 3 日 18 时许,被告人于某在本市海淀区莲花桥东南角辅路路边,因行车纠纷与被害人王某(女,32 岁)、吴某(男,35 岁)发生口角,遂用拳脚将被害人王某、吴某打伤,致被害人王某左上中切牙冠根折,右上中切牙半脱位、头面部软组织损伤,经鉴定为轻伤二级;致被害人吴某头面部软组织伤、头皮血肿、右侧胸肋部软组织伤,经鉴定为轻微伤,同年 3 月 21,被告人于某被公安机关抓获归案,后如实供述了上述犯罪事实。2015 年 5 月 28 日,被告人于某在家属的帮助下赔偿被害人王某、吴某人民币 11 万元,双方达成和解。	[234,67]	6

(下转第 13 页)