

基于改进孪生网络的相似案件匹配算法

费志伟¹, 艾中良², 张可¹, 刘忠麟¹, 唐然¹

(1. 华北计算技术研究所, 北京 100083;

2. 中国司法大数据研究院, 北京 100043)

摘要:类案检索是人工智能在司法领域中应用研究的热点之一。针对现有类案检索系统中存在的相似案件检索准确率低、相似案件匹配模型复杂等问题,设计了一种基于改进孪生网络的相似案件匹配算法。该算法使用改进的孪生网络对输入案件事实的相似性进行评估,通过引入注意力机制和 Dropout 层实现了用于计算相似性的非对称的孪生网络结构。提出基于启发式和自增广的训练数据增强策略,扩充训练数据集,进一步提高模型预测准确率和对输入顺序的鲁棒性。在公开数据集 CAIL2019-SCM-Large 上进行实验验证,并对比了其他方法。实验结果表明,该方法在该数据集上准确率达到 72.59%,相比于现有方法有很大的提升,证明了该方法的有效性。

关键词:类案匹配;孪生网络;注意力机制;文本数据增强;文本分类

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2022)0152-06

Similar Case Matching Algorithm Based on Improved Siamese Network

FEI Zhi-wei¹, AI Zhong-liang², ZHANG Ke¹, LIU Zhong-lin¹, TANG Ran¹

(1. North China Institute of Computing Technology, Beijing 100083, China;

2. China Justice Big Data Institute, Beijing 100043, China)

Abstract: Similar cases retrieval is one of the hotspots in the application of artificial intelligence in the judicial field. Aiming at the problems of low accuracy of similar case retrieval and complex matching model of similar cases in the existing similar case retrieval system, a similar case matching algorithm based on improved siamese network is designed. The algorithm uses the improved siamese network to evaluate the similarity of the input case facts, and realizes the asymmetric siamese network structure by introducing the attention mechanism and the Dropout layer. A training data augment strategy based on heuristics and self-augmentation is proposed, which realizes the expansion of training data set for asymmetric siamese network and further improve the accuracy and robustness to the input sequence. Experiments were performed on the public data set CAIL2019-SCM-Large, and other methods were compared. The experimental results show that the accuracy of this method on this data set is 72.59%. The proposed model outperforms other existing methods on this data set, which proves the effectiveness of this method.

Key words: similar case matching; siamese network; attention mechanism; text data augmentation; text classification

0 引言

司法公正是社会关注的焦点,在法官审理案件过程中,相同或者类似的案件能否得到相同或类似的判决结果是人们普遍关注的内容。“类案类判”是我国司法领域追求的目标,也是提升司法公信力,统一裁判标准,避免司法裁判不公的重要举措。随着我国司法信息化的推进,尝试将人工智能技术与类案检索相结合,建立类案检索系统是目前最为热门的一项研究。类案检索系统通过主动或被动的方式为法官提供与正

在处理的案件相似甚至相同的案件,来启发、拓展法官判决思路、帮助法官实现正确裁判,提高法官办案效率,法官对类案检索系统有着强烈的需求和期待。

目前已有部分法院上线了类案检索系统,2018年1月,在最高法院上线了“类案智能推送系统”,据介绍,该系统从“案件性质、案情特征、争议焦点、法律适用”四个方面,覆盖全部1330个案由,通过机器学习构建出超过10万个维度的特征体系;在精准度方面,做到全案由文书数据整体搜索推送准确率达到

收稿日期:2021-10-25

基金项目:国家重点研发计划(2018YFC0832306,2018YFC0831203,2018YFC0831206)

作者简介:费志伟(1996-),男,研究生,CCF会员(C6428G),研究方向为自然语言处理、文本生成;艾中良,硕士,研究员,研究方向为大数据技术、人工智能技术。

63.7%,民事、刑事 Top10 类型的准确率达到 85.5%,其中检索全案由整体准确率 61.6%,热门类型案由整体准确率 81.8%^[1]。其他地方法院也相继上线试点了一些类案检索系统,虽然类案检索系统在部分法院上线试点运行,但类案检索仍面临着诸多难点。目前类案匹配准确率低,模型复杂,需要更好的类案检索机制和匹配算法。

针对这些问题,该文提出并实现了一个法律文本相似案件匹配算法,该算法能够根据案件事实部分,判断需要比较的案件与该案件事实是否相似。该算法在孪生网络框架的基础上,通过在获取的句特征后加入注意力机制和 dropout 层,对句特征做增强,同时提出一种训练数据增强策略,充分训练该网络,让其对输入顺序不敏感。在 CAIL-2019-Large 数据集^[2]上做测试,相比于传统模型有很大的提升。该算法能在较少的数据集上实现比较好的匹配准确率,对之后的类案检索系统设计有很大的参考价值。

1 相关工作

类案检索是对给定的查询案例,给出案件事实或者关键案情要素和构成要件与之相近的案件^[3]。目前类案检索方法主要通过文本语义相似性计算来对两个案件进行比较。文本语义相似性计算是通过机器学习的方式识别文本语义内容,比较两个文本在语义上是否相似。其作为自然语言处理中的基础问题,在问答、机器翻译、信息检索等领域被广泛应用。常见的直观方法是将文本转化为基于词的向量空间,如 TF-IDF^[4]、词袋模型,通过词频信息,比较在这种表示下的文本相似性^[5]。但是法律文本在结构上和词汇上有着相似之处,法官关注的内容与通过机器学习方式自动识别出的内容存在一定的偏差,法官根据自身的法律知识需要精准类案检索系统,这种方法并未取得很好的效果。在此基础上改进的方法是案件要素匹配技术,案件要素匹配技术主要通过对案件贴标签,识别案件的法律标签,如上述高院试点的类案检索系统,通过识别案件中的法律标签如“自首”、“坦白”、“多次盗窃”等,在对案件进行相似性计算时通过比较法律标签来识别两个案件的相似性。相比于 TF-IDF、词袋模型这些方法,该方法更贴近司法实际,能够为法官提供明确的指导,但自动或人工的方式梳理法律标签增加了模型的复杂性。王君泽等^[6]将案件相似性转化为案情内容各自所含语句之间的相似度计算,依据词项重要性、语句内容和数目综合对一起案件进行考察,得到其于其他案件的相似性。其他有基于文本整体构建向量和主题模型^[7]等一些方法。

随着深度学习技术的发展和相关中文数据集的出

现,出现了一些基于深度神经网络的相似文本匹配技术。深度学习技术能够自动的提取文本特征,简化了人工提取特征和梳理关键词等特征工程带来的复杂的操作。李兰君等^[8]提出一种基于改进孪生网络的相似法律案例检索研究,利用层级注意力机制的孪生网络模型计算法律案例文档相似性,在自建的小规模数据集上测试,评价指标采用皮尔孙相关系数和均方误差,达到皮尔孙相关系数 0.97 和均方误差 0.006 效果。2019 年中国司法大数据研究院和北京大学、清华大学等高校举办了“法研杯”比赛,在评测任务中公开了中文相似案件匹配数据集 CAIL-2019-SCM,该数据集由民事借贷纠纷组成,在构建数据集时通过标注法律要素的方式来判断类案。在比赛中基于孪生网络的相似文本匹配模型和基于 Triplet 网络^[9]的深度学习方法被广泛使用。

从比赛结果可以看到,目前在该数据集上最优的模型采用多模型融合的方式,通过 5 个改进的 Triplet 模型进行融合,损失函数使用 Triplet loss^[10],并参考合同法、担保法、婚姻法及相关司法解释,总结并使用了四个特征作为数据的扩充,最终在当年比赛中取得第一名。

2 相似案件匹配算法

2.1 任务定义

相似案件匹配主要计算案件事实之间的相似性,该文构建的文本相似性计算案例,将三个案件事实内容组成三元组 (A, B, C) 作为输入,其中案件 A 是查询案件,候选集为案件 B, C 。对候选集中案件 B, C ,其中一个为正样例,另一个是负样例。使用这种方式构建数据集主要关注相似性上,而非案件查询排序上。三元组的形式化表示为 $a = \{a_1, a_2, \dots, a_{l_a}\}$, $b = \{b_1, b_2, \dots, b_{l_b}\}$ 和 $c = \{c_1, c_2, \dots, c_{l_c}\}$ 。其中 a_i, b_j, c_k 是相似案件中的字符, l_a, l_b, l_c 表示案件文本 a, b, c 的长度。相似案件匹配的目标就是判断案件事实 A , 在两个候选集 B, C 中与哪一个候选案情更加相似。任务的目标是预测标签 $y \in \{0, 1\}$, 当 $y = 0$ 时, $\text{sim}(a, b) > \text{sim}(a, c)$, 或者当 $y = 1$ 时, $\text{sim}(a, b) < \text{sim}(a, c)$ 。

2.2 算法整体流程

该文将三个文本处理成文本对,通过将待比较的文本与给定文本拼接起来得到文本对,然后对文本对进行分类得到相似的案件。通过交换拼接顺序和增加拼接方式对训练数据做文本增强。在孪生网络模型中通过引入注意力机制和 Dropout 层对文本特征进行增强,然后通过做差获取两个文本对特征上的差异,最后通过全连接层得到最终的结果,算法整体流程如图 1 所示。

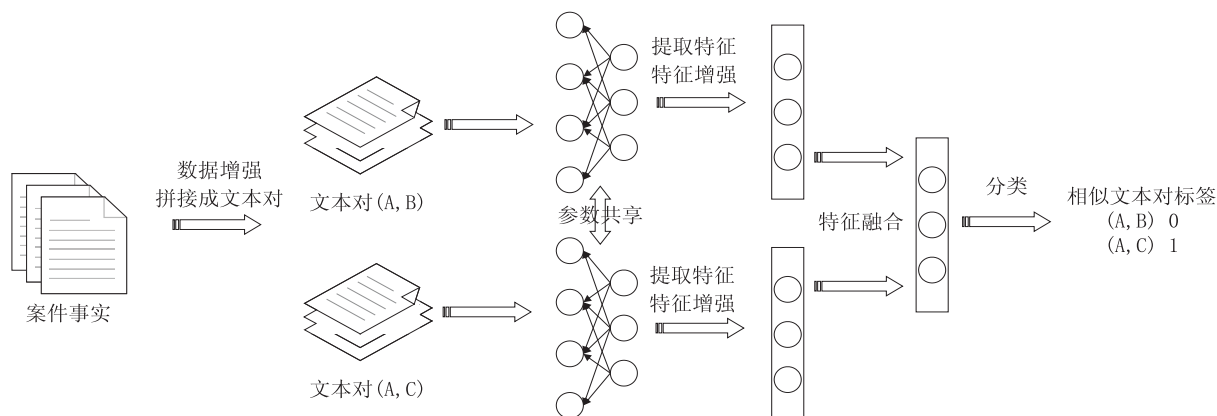


图1 算法流程

2.3 改进孪生网络算法

孪生网络^[11]由两个结构相同,且权重共享的神经网络耦合而成。将样本通过神经网络映射到特征空间,然后通过度量函数计算距离来比较两个样本是否相似。在文本领域,Neculoiu 等^[12]提出基于多层双向长短期记忆网络的孪生网络来计算文本之间的相似性。通过多层双向长短期记忆网络提取文本特征,然后通过全连接层将特征降维,最后通过能量函数度量降维后的特征的相似性。现有的孪生网络架构大多基于以上架构,或者改进特征提取部分,将多层长短期记忆网络替换成 GRU^[13]、CNN^[14]、BERT^[15]等模型,模

型的输入为 (x_1, x_2, y) , 其中 x_1, x_2 为需要比较相似性的文本, $y \in \{0, 1\}$ 为标签, 相似的文本记为 0, 不相似文本记为 1。

现有的基于孪生网络的相似文本匹配方法对三元组文本 (A, B, C) 需要比较 (A, B) 或 (A, C) 之间的相似性, 每次只能利用三元组中的两个, 未能很好的利用到全部的信息。该文提出的方案将两个文本拼接后送入孪生模型中得到对应的文本表示, 该方法能同时利用到三元组中三个文本的信息。改进后的孪生网络结构如图 2 所示。

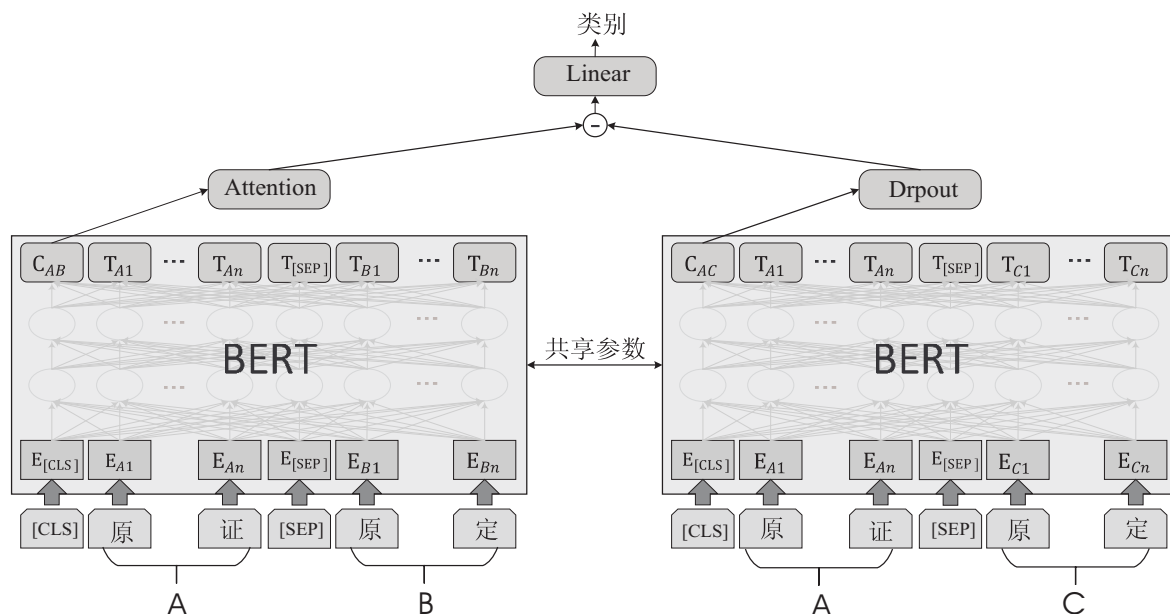


图2 改进孪生网络架构

该算法打破了孪生网络设计上的对称性,采用非对称结构设计,算法细节如下:

将文本送入中文预训练语言模型中,获取最后一层池化后的隐层参数得到特征 h_{AB} , h_{AC} 。 $h_{AC} \in \mathbb{R}^{d_h}$, 其中 d_h 为隐藏层维度。

$$h_{AB} = \text{Encoder}(D_{AB})$$

$$h_{AC} = \text{Encoder}(D_{AC})$$

将特征 h_{AB} 通过注意力机制^[16]进行增强,增强对判断文本对相似性的特征。注意力机制选取了自注意力机制。

$$a_{AB} = \text{Attention}(h_{AB})$$

特征 h_{AC} 通过 Dropout 层^[17]做增强。Dropout 层通过随机丢弃一部分神经元同时丢弃这些神经元对应的连接边,来避免过拟合。在此处加入 Dropout 层相当

于给文本特征加入噪声,做特征层面的数据增强,提升模型的鲁棒性。

$$d_{AC} = \text{Dropout}(h_{AC})$$

最后将增强的文本特征和 dropout 后的特征做差,获得两个文本的特征表示,再通过全连接层和 softmax 激活函数进行相似文本对分类。

$$m = a_{AB} - d_{AC}$$

$$y = \text{softmax}(mW + b)$$

其中, y 为标签, $y \in \{0, 1\}$ 。 $W \in R^{d_1 \times d_2}$ 为一个待学习的矩阵。

损失函数选取了交叉熵损失函数。

2.4 数据增强方法

上述模型中,加入注意力机制和 dropout 层使得该神经网络结构不对称,为了提高非对称孪生网络模型对输入文本对顺序上的鲁棒性,让其不会因交换输入顺序而导致结果不同,在数据上对训练数据集进行扩充来充分的训练模型。将法律文本三元组记为 (A, B, C) ,其中拼接得到的法律文本对记为 $(A, B)(A, C)$ 。将 (A, B) 相似对应到标签 0, (A, C) 相似对应到标签 1。将 $\text{sim}(\cdot)$ 记作相似性度量函数。

模型对于交换性应该是不敏感,交换文本对的输入不能改变文本的标签。其次希望文本 (A, B) 之间尽量相似, (A, C) 之间尽量不相似,通过增加 (A, B) 和 (B, C) 之间的相似性和 (A, C) 和 (C, B) 之间的相似性来扩充数据。假设了 $\text{sim}(A, B) < \text{sim}(B, C)$ 和 $\text{sim}(A, C) > \text{sim}(B, C)$ 。结合上述交换律,扩充数据对和标签信息如表 1 所示。

表 1 启发式扩充数据标签

启发式增广	标签
$(A, B), (A, C)$	0
$(A, C), (A, B)$	1
$(B, A), (B, C)$	0
$(B, C), (B, A)$	1
$(C, B), (C, A)$	0
$(C, A), (C, B)$	1

之后通过使用 A 数据和 C 数据通过重复的方式来对相似文本数据对数据集进行扩充。通过将文本 A 组合成 (A, A) 形式来扩充相似文本内容。

表 2 自增广扩充数据标签

自增广	标签
$(A, A), (A, C)$	0
$(A, C), (A, A)$	1
$(C, C), (C, A)$	0
$(C, A), (C, C)$	1

通过上述方法进行数据扩充后再去除重复的部分,作为扩充后的训练数据。

3 实验结果与分析

3.1 评价指标

测试采用最新的 CAIL2019-SCM-Large 数据集,相比之前发布的版本,该数据集划分为 5 102 条训练数据、1 500 条验证数据和 1 536 条测试数据。数据内容为民间借贷案件,采用三元组的方式给出三个案件的事实描述。该数据集给出三元组 (A, B, C) 和一个标签的形式,通过给出文本与 A 文本更相似的文本标签取代之前默认保证 A 与 B 之间的相似度大于 A 与 C 之间的相似度。

测试通过判断给定的两个文本与待比较文本给出更相似的文本标签,对标签进行判断,如果正确记为 1,为预测正确的三元组不得分,统计所有正确的标签个数除以整个测试集数得到最终准确率。

$$\text{Acc} = \frac{T}{S}$$

其中, T 为预测正确的三元组个数, S 为总共三元组个数。

3.2 实验参数

在训练时采用 warm-up 策略,优化函数选择 AdamW 优化器^[18],训练相关超参数如表 3 所示。

表 3 模型超参数设置

超参数名	数值
warm-up rate	0.1
learning-rate	$2e^{-5}$
dropout rate	0.3
batch size	16

3.3 实验结果分析

在预训练模型上,该文测试了科大讯飞公开的中文预训练语言模型和清华大学公开的司法预训练语言模型,科大讯飞的中文预训练语言模型^[19]有三种 BERT-wwm、BERT-wwm-ext 和 RoBERT-wwm-ext,均采用全词遮蔽策略,在中文维基百科(包括简体和繁体)上进行预训练,其中 ext 使用其他百科、新闻、问答等数据扩充了预训练数据。清华大学公开的司法预训练模型中,使用了裁判文书网上民事案件数据做预训练。这些模型的最大输入长度都是 512 字符,输出的隐藏层为 768 维。

实验对比了多个预训练语言模型,结果如表 4 所示。从实验数据能够看到 bert-wwm-ext 取得了最佳的成绩,三个预训练模型差距并不大,说明在该问题上预训练模型并没有起到关键的作用。

表 4 预训练模型和准确率

Model	Test acc
BERT-wwm	71.68
BERT-wwm-ext	72.59
BERT-ms	71.55
RoBERTa-wwm-ext	72.20

以上模型的准确率来自钟皓曦等^[2]论文中的数

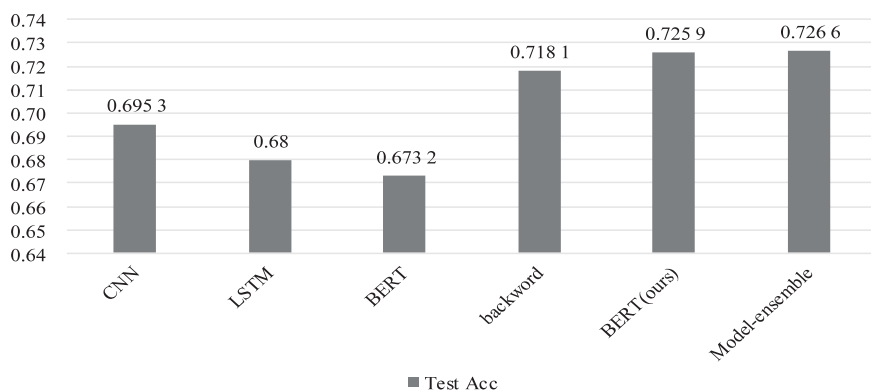


图 3 多个模型测试集准确率

该文提出的模型与融合案件要素信息的集成模型之间准确率仅相差 0.07%, 及预测对的案件数仅差 1 条, 相比于 backword 队提出的模型提升了 0.78%, 达到目前单个模型最优水平。由此可以看出, 目前多模型融合以及根据法律知识提取要素的方式并没有起到很好的效果, 现有的方法在相似案件匹配数据集在整体上效果并不好, 该问题还有很大的提升空间。

3.4 消融实验

该文比较了修改网络结构和不使用数据增强的几种方式, 在模型结构上比较了在孪生网络结构后取相同的特征增强层, 同为注意力层或同为 dropout 层的结果。在数据增强上比较了不使用数据增强, 或者只使用一种数据增强的结果, 实验结果如表 5 所示, 实验默认在提出的模型和策略作为基础, 表中 w/o 表示不包含此方法。

表 5 消融实验结果

方法	测试准确率/%
相同特征提取层(注意力层)	71.22
相同特征提取层(dropout 层)	71.03
文中模型 w/o 数据增强	69.85
文中模型 w/o 自增广	72.26
文中模型 w/o 启发式增广	71.55

从实验结果看到在特征增强层取相同的结构并不能带来很好的提升, 相同的特征提取层如加入 dropout 层会降低准确率。在不使用数据增强时, 本模型的准确率大幅下降, 数据增强结合文中提出的特征增强方法能有效提高准确率。在测试自增广和启发式增广

据, 从多个模型的测试结果可以看出, 基于 CNN 和 LSTM 的模型效果好于 BERT 模型, 说明模型输入文本长度对最终结果有很大的影响。BERT512 字符的输入长度限制了模型获取更多的文本内容, 从而获得对分类结果更有效的特征信息。改进后的模型远优于基于孪生网络的 BERT 模型, 说明上述改进方案简单有效的提高了模型的性能。

时, 两种方法均能提高准确率, 启发式增广提高的准确率明显, 这与文中的数据集有一定的关系, 文中数据集文本长度相加超出 BERT 能输入的最长文本内容, 启发式增广在该数据集上能使模型获得更多的文本内容上的信息和文本对问的关系。在不使用自增广的情形下, 准确率下降并不明显, 说明自增广测量能小幅度的提升模型的准确率, 在提高模型准确率上, 启发式数据增广作用更强。

4 结束语

实现类案类判是司法实务中人们比较关心的一方面, 提出一种改进孪生网络算法和一种数据增强方式。该方法使用单个模型, 在 CAIL-2019-SCM 测试数据集上取得目前单个模型最优水平。相比于目前最佳的多模型融合和通过人工梳理特征进行增强的策略, 该方法是简单有效的, 并与多模型融合算法的结果相差极小, 进一步说明该模型的优越性。目前在相似案件匹配数据集上, 现有方法还有很大的提升空间, 需要进一步思考如何通过引入法官判断类案关注的信息来进一步对类案匹配进行细化, 构建高效的、可解释的类案匹配模型。

参考文献:

- [1] 左卫民. 如何通过人工智能实现类案类判[J]. 中国法律评论, 2018, 20(2): 39-45.
- [2] XIAO C, ZHONG H, GUO Z, et al. CAIL2019-SCM: a dataset of similar case matching in legal domain[EB/OL]. (2019-11-25) [2021-05-31]. <https://arxiv.org/abs/>

- 1911.08962.
- [3] MA Y, SHAO Y, WU Y, et al. LeCaRD: a legal case retrieval dataset for Chinese law system[C]//SIGIR '21: the 44th international ACM SIGIR conference on research and development in information retrieval. NY: ACM, 2021.
- [4] 黄承慧, 印 鉴, 侯 昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度度量方法[J]. 计算机学报, 2011(5): 856-864.
- [5] 王春柳, 杨永辉, 邓 霏, 等. 文本相似度计算方法研究综述[J]. 情报科学, 2019, 37(3): 158-168.
- [6] 王君泽, 马洪晶, 张 毅, 等. 裁判文书类案推送中的案情相似度计算模型研究[J]. 计算机工程与科学, 2019, 41(12): 2193-2201.
- [7] VENKATESH R K. Legal documents clustering and summarization using hierarchical latent Dirichlet allocation [J]. IAES International Journal of Artificial Intelligence, 2013, 2(1): 27-35.
- [8] 李兰君, 周俊生, 顾颜慧, 等. 基于改进孪生网络结构的相似法律案例检索研究[J]. 北京大学学报: 自然科学版, 2019, 55(1): 87-93.
- [9] HOFFER E, AILON N. Deep metric learning using triplet network [C]//International workshop on similarity - based pattern recognition. [s. l.]: Springer, 2015: 84-92.
- [10] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: a unified embedding for face recognition and clustering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Boston: IEEE, 2015: 815-823.
- [11] BROMLEY J, GUYON I, LECUN Y, et al. Signature verification using a "siamese" time delay neural network[J]. Advances in Neural Information Processing Systems, 1993, 6: 737-744. 1
- [12] NECULOIU P, VERSTEEGH M, ROTARU M. Learning text similarity with siamese recurrent networks[C]//Proceedings of the 1st workshop on representation learning for NLP. Berlin: ACL, 2016: 148-157.
- [13] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. (2014-12-11) [2021-05-31]. <https://arxiv.org/abs/1412.3555>.
- [14] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha: ACL, 2014: 1746-1751.
- [15] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2018-10-11) [2021-05-31]. <https://arxiv.org/abs/1810.04805>.
- [16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-12-06) [2021-12-08]. <https://arxiv.org/abs/1706.03762>.
- [17] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
- [18] LOSCHILOV I, HUTTER F. Fixing weight decay regularization in Adam[EB/OL]. (2017-11-15) [2021-12-08]. <https://arxiv.org/abs/1711.05101>.
- [19] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for chinese bert[EB/OL]. (2019-10-29) [2021-12-08]. <https://arxiv.org/abs/1906.08101>.
- +++++
- (上接第 151 页)
- on remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015, 8(8): 4206-4222.
- [9] WU Xi, SHI Zhenwei. Scene aggregation network for cloud detection on remote sensing imagery[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 10: 1-5.
- [10] 时红伟. 一种面向用户任务需求的遥感图像质量标准 2NIIRS[J]. 航天返回与遥感, 2003, 24(3): 30-35.
- [11] 刘 迪, 李迎春. 基于深度学习和人眼视觉系统的遥感图像质量评价[J]. 激光与光电子学进展, 2019, 56(6): 153-161.
- [12] 产晓冰, 周川杰, 吕 原. 基于应用的可见光遥感图像质量评价方法研究[J]. 无线电工程, 2013, 43(11): 50-52.
- [13] 冈萨雷斯. 数字图像处理[M]. 第 4 版. 北京: 电子工业出版社, 2020.
- [14] 刘成龙. MATLAB 图像处理[M]. 北京: 清华大学出版社, 2017.
- [15] KLETTE R, ROSENFELD A. Digital geometry - geometric methods for digital picture analysis[M]. San Francisco, CA: Morgan Kaufmann, 2004.