第 11 卷 第 2 期
2022 年 3 月

集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY

Vol. 11 No. 2
Mar. 2022

引文格式：
周献杭, 申妍燕. 基于多粒度语义交互的无监督法律裁判文书检索 [J]. 集成技术, 2022, 11(2): 55-66.
Zhou XH, Shen YY. Unsupervised legal case retrieval based on multi-granularity semantic-aware interaction [J]. Journal of Integration Technology, 2022, 11(2): 55-66.

基于多粒度语义交互的无监督法律裁判文书检索

周献杭 申妍燕*

（中国科学院深圳先进技术研究院 深圳 518055）

摘 要 随着法律文书数据越来越多，信息过载问题日益严重，快速且准确地在海量法律文书中进行检索显得非常必要。法律文本作为一种特殊的文本形式，具有篇幅较长、结构复杂、专业性强等特点，传统基于关键字的文本检索方法不能满足用户查询法律信息的需求，容易出现答非所问、检索不全等问题。此外，基于语义的文本检索方法，大多依赖于对含有大量标注数据的法律文本进行有监督学习，而法律文本数据的人工标注则严重依赖专家知识，导致其需要高昂的人力成本。该文提出一种基于无监督学习的法律文书检索模型，分别从法律概念、词语和词组 3 个方面进行多粒度无监督文本匹配，避免了没有训练数据导致的冷启动问题。在法律裁判文书数据集上进行检索实验的结果表明，与基准模型相比，该模型在 MAP、MRR 和 NDCG@10 指标上均有显著提升，取得了优秀的检索效果，具有有效性和先进性。

关键词 无监督学习；文本检索；法律文书检索；多粒度语义交互
中图分类号 TP 399 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20210413001

Unsupervised Legal Case Retrieval Based on Multi-granularity Semantic-Aware Interaction

ZHOU Xianhang SHEN Yanyan*

(*Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)

**Corresponding Author*: *yy.shen@siat.ac.cn*

Abstract With the ever-increasing size of legal cases in China, relevant legal case retrieval given a user query has attracted considerable attention. Conventional keyword-based retrieval systems look for matching cases that contain one or more words specified by the user. However, keyword searching is sharply focused on finding the exact terms specified in the query, making the retrieval systems miss many relevant documents. On the other hand, semantic-aware information retrieval methods usually rely heavily

收稿日期：2021-04-13 **修回日期：**2021-06-15
基金项目：法律人工智能联合实验室项目(Y9Z028)
作者简介：周献杭，硕士研究生，研究方向为自然语言处理-信息检索；申妍燕(通讯作者)，副研究员，研究方向包括无人机通信、移动边缘计算以及深度强化学习算法在无线通信中的应用，E-mail：yy.shen@siat.ac.cn。

on labeled training data. Nevertheless, obtaining rich annotated data is a time-consuming and expensive process, creating a substantial barrier for applying supervised methods to legal case retrieval. This paper proposes an unsupervised, semantic-aware legal case retrieval method based on multi-granularity multi-granularity semantic interaction. Specifically, leveraging legal concept-level features, phrase-level features, and word-level features to comprehensively explore the semantic-aware interaction between the query and each legal case. Experimental results on a real-life legal case retrieval corpus demonstrate that this method has a significant improvement in MAP, MRR and NDCG@10 indicators and substantially outperforms the baseline methods.

Keywords unsupervised learning; text retrieval; legal case retrieval; multi-granularity semantic interaction

Funding This work is supported by SIAT-DELI Artificial Intelligence and Law Lab (Y9Z028)

1 引 言

近年来，司法部持续加强法律服务智能化建设，深化“数字法治、智慧司法”建设，旨在建设覆盖行政立法、行政执法、刑事执行、公共法律服务四大职能的“大平台、大系统、大数据”。随着互联网技术的飞速发展，以及裁判文书网等司法公开措施的推进，人们可以获得越来越多的裁判文书，但信息过载问题也日益严重，快速且准确地对海量法律文书进行检索显得非常必要。如，法官通过检索相似案例来获取办案参考；律师援引相似案件来作为论证依据等。

法律裁判文书检索(Legal Case Retrieval)[1]，是指从历史裁判文书集合中抽取与用户查询相关的裁判文书的任务。具体地说，文本检索模型通过计算查询语句(或文档)与每个待检索裁判文书之间的相似度，然后根据相似度大小对裁判文书进行排序，从而返回满足用户需求的裁判文书集。法律文本作为一种特殊的文本形式，具有篇幅较长、结构复杂、专业性强等特点，传统的基于关键字的文本检索方法已经不能满足用户查询法律信息的需求。基于关键字的检索方法主要是对用户查询关键字进行匹配，虽然能够快速查找到所需的信息，但该类方法容易出现查准率低、检索不全等问题。此外，新用户很难通过识别合适的关键词来准确检索法律裁判文书。

随着深度学习的快速发展，学者们对语义检索进行了深入的研究和探索。语义检索能够更加准确地通过用户输入的查询信息理解其真正意图，从而具有更高的查全率和查准率[2]。然而，基于语义的文本检索方法大多依赖于有大量标注数据的有监督学习方法，并且法律文本数据的人工标注过程严重依赖专家知识，费时费力，这无疑给研究工作带来巨大的人力成本。

为了更好地解决上述问题，本文提出一种基于无监督学习的法律裁判文书检索模型，分别从法律概念级别、词语级别和词组级别 3 个方面进行多粒度文本匹配，避免了没有训练数据导致的冷启动问题。具体地说，该模型利用词语级别的交互特征来捕获用户输入查询和候选文档的关键词信息；利用词组级别的交互特征来更加灵活地捕获特定表达信息；利用官方法律案由级别信息和提取的案由数据构建法律知识词典，并通过该词典捕获用户查询和文档之间的法律概念级别匹配信息。然后深度融合法律概念级别、词语级别和词组级别的匹配信号，得到最终的排序结果。实验结果表明，与基准模型相比，本文模型在法律裁判文书检索

上取得了更好的效果，具有有效性和先进性。

2　检索算法研究现状

信息检索(Information Retrieval，IR)[3]任务是将用户的检索查询内容与数据库中大量的文档集合根据特定的条件进行相关性建模，计算检索语句与文档间的相关程度并召回，再对召回的文档根据用户需求重新排序后返回给用户。在过去几十年的不断探索实践中，已经相继提出许多不同的检索模型，包括布尔模型[4]、向量空间模型[5]和概率模型[6]等传统检索模型，以及统计语言模型[7]和排序学习模型[8]。这些检索模型成为了信息检索技术研究的基础。其中，基于排序学习的检索模型已经在许多应用中取得了巨大的成功，至今仍被学者不断研究改进，并广泛应用于商业检索系统。近年来，大数据研究掀起的机器学习热潮，在语音识别、计算机视觉和自然语言处理任务中取得了令人激动的成果，加速了信息检索技术的发展。目前，信息检索的研究主要分为两个方向——基于表示学习的检索模型和基于匹配函数学习的检索模型[9]。

2.1　传统检索模型

1972 年，Jones[10]提出了 TF-IDF 模型，该模型将文档和检索语句转换为特征向量，并利用加权技术评估一个字词在文档集合中的重要程度。这个模型在 20 世纪 80 年代一直是科研人员的研究重点，直到现在，基于 TF-IDF 的衍生模型仍然被搜索引擎采用。1992 年，Robertson 等[11]提出了基于概率的 BM25 模型，该模型在传统 TF-IDF 模型的基础上增加了可调节的参数，并且将文档长度作为惩罚因子，使检索模型具有更高的实用性。该模型在许多信息检索任务中表现优异，成为重要的基准算法。

传统检索模型主要通过提取词条的特征信息，利用集合论知识、代数模型和概率模型方法进行检索，不需要标注数据的支持和复杂的模型结构，属于无监督检索模型范畴，可以有效解决没有训练数据导致的冷启动问题。传统检索模型依靠特征提取，但原始文档中的噪声信息和进行句法分析时累积的匹配误差会引起错误的召回，从而影响最终的检索效果。此外，传统检索模型会忽略语料库外词语的特征信息，同时使用孤立的关键词特征脱离了上下文语言环境，难以还原语义信息，从而影响模型最终的检索效果。

2.2　基于表示学习的模型

基于表示学习的模型会分别学习输入检索语句和数据库中文档的高层向量表示，通过构造评估函数，来得到检索语句和文档集合的匹配得分。2013 年，微软 Huang 等[12]提出 DSSM 模型，这是首个成功地将表示学习应用在文本匹配任务的模型，其采用有监督训练方式，不需对文档集合进行特征标注工程，且利用预先计算文本向量可大幅降低在线计算的消耗，该模型为深度语义匹配模型的鼻祖。2014 年，微软 Shen 等[13]提出的 CNN-DSSM，以及 2016 年，Palangi 等[14]提出的基于循环神经网络的 LSTM-RNN 模型，都对 DSSM 模型进行了优化，但两者仍需要大量的训练数据及 GPU 支持，导致许多业务难以开展。

2.3　基于匹配函数学习的模型

基于匹配函数学习的模型，主要围绕构建交互函数以获得检索语句和文档中词语的详细交互信号。2014 年，Hu 等[15]在 NIPS 上提出了 ARCII 模型，该模型通过对检索语句和文档中的单词进行 N-Gram 的卷积提取词序信息，然后对各自卷积后得到的词向量按对进行计算，从而得到一个匹配度矩阵。2016 年，Guo 等[16]提出了 DRMM 模型，该模型将检索语句和文档分别表示为由 M 个和 N 个词语组成的向量。将检索语句和文档逐一比对计算相似度，以直方图的形式进行分桶计算代替池化操作，可以更好地区分相似和完全匹

配，并保留原始信息。基于匹配函数学习的模型规避了对长文本进行准确编码的困难，但构造交互矩阵难以进行在线计算，效率有限。

2.4 面向法律领域的匹配模型

在法律文本匹配领域[17]，可利用词嵌入的方式在文本和向量之间建立连接[18]。有的学者在基于符号的方法上引入可解释的标签[19]，从而在法律文件中的符号之间进行推理。Sugathadasa 等[20]将法律案件转移至向量空间，从而进行文档检索任务。Chalkidis 和 Kampas[21]主要围绕将现有的词嵌入方法（如 word2vec[22]）应用于法律语料库。Tran 等[23]设计了一种编码摘要模型，该模型将给定法律文档编码到连续的向量空间中，同时嵌入文档的摘要属性。法律领域的匹配模型仍在快速发展，为了获取法律领域专业词汇的表述，可以在词嵌入中捕获法律知识，Liu 等[24]将知识图谱结构化信息引入神经网络检索模型中。Zhang 等[25]利用图嵌入技术，使神经网络检索模型能够利用图的结构化数据自动进行特征提取，这不仅有助于克服点击数据的长尾问题，而且可以通过结合外部信息来改善搜索结果。由于法律领域包含许多规则和知识，将知识进行建模对于法律文本匹配同样至关重要[26]。

与通用领域的检索模型相比，法律领域的检索模型面向大量结构化的法律文献、法律法规和案件文书，若采用传统的关键字检索模型，则无法得到关键词之间的内在联系，因此，检索文档的召回率和排序的准确率难以得到保证。由于法律信息包含司法专家长期实践的经验以及法律领域专业知识汇总形成的专业术语和复杂的结构表达，所以想要搜索出检索语句和文档之间更深层的含义，在获得高召回率的同时，提高排序的准确率以及增加检索结果的可用性，仍然有许多障碍[26]。因此，在前期没有标注数据的前提下，根据用户输入的自然检索语句，构建基于多粒度语义交互的无监督法律裁判文书检索模型，优化法律领域的文献检索，提升排序准确率具有现实意义。

3 无监督法律文书检索模型

多粒度语义交互的无监督法律裁判文书检索模型（Unsupervised Legal Case Retrieval based on Multi-granularity Semantic-aware Interaction，ULRM）的输入为用户的检索语句，模型针对检索语句分别从词语级别、词组级别、法律概念级别进行信息匹配，综合 3 个模块的打分结果来决定召回文书的最终排名顺序，随后输出精排后的裁判文书并反馈给用户。利用词语级别的交互特征来捕获用户输入查询和候选文档的关键词信息；利用词组级别的交互特征来更加灵活地捕获法律领域的特定表达；利用官方法律案由级别信息和提取的案由数据构建法律知识词典，从而更准确地匹配用户的需求，提升检索效果。ULRM 模型结构如图 1 所示。

3.1 词语级别匹配

对于用户以自然语言形式输入检索语句，系统地通过分词和去停用词预处理将检索语句转换成关键字序列，并利用词嵌入方式将检索关键字序列转换成矩阵 $\boldsymbol{Q}_{n\times k}$。对于数据库中的法律文书进行相同操作，将其转换成矩阵 $\boldsymbol{D}_{m\times k}$。

$$\boldsymbol{Q}_{n\times k}=\left[\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_{n-1},\boldsymbol{q}_n\right] \tag{1}$$

$$\boldsymbol{D}_{m\times k}=\left[\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_{m-1},\boldsymbol{d}_m\right] \tag{2}$$

其中，n 为用户输入的检索语句经分词后的关键词个数；m 为数据库中裁判文书经分词后的关键词个数；k 为关键词转换成词嵌入后向量的维数；$\boldsymbol{q}_i$ 为检索语句经预处理后序列中第 i 个词向量，维度为 k；$\boldsymbol{d}_j$ 为裁判文书经预处理后序列中第 j 个词向量，维度为 k。

通过分词预处理和词嵌入得到词向量后，利用余弦距离来衡量检索语句和裁判文书中词向量的相关性信息，再通过计算余弦相似度得到两者

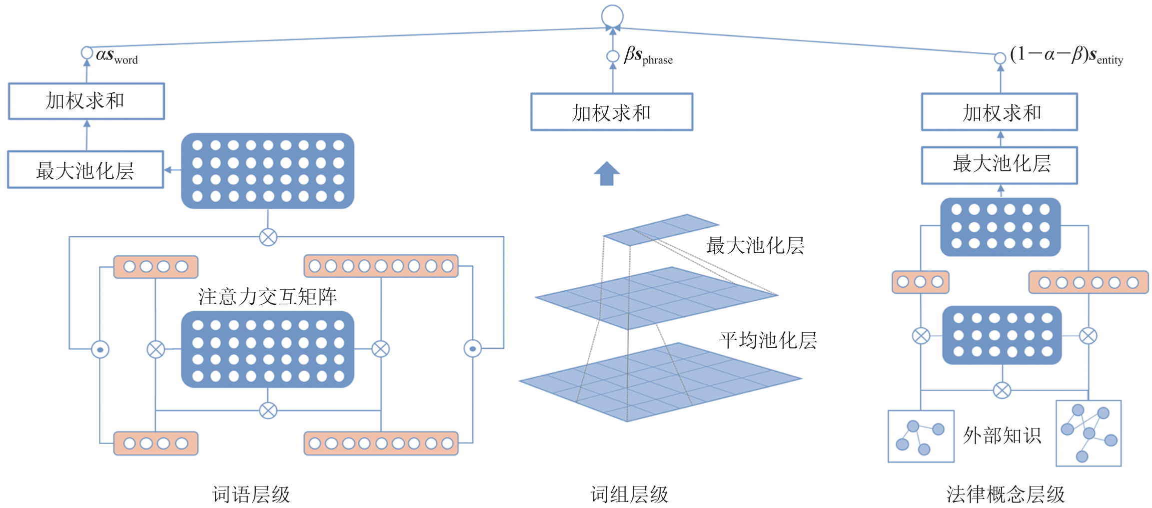


图 1　ULRM 模型结构图

Fig. 1　The architecture of the ULRM

的相关性矩阵 $\boldsymbol{M}_{n\times m}$。

$$M_{i,j}=\cos\left(\boldsymbol{v}_i^q,\boldsymbol{v}_j^d\right) \tag{3}$$

其中，$M_{i,j}$为检索语句中的第 i 个词向量与文书中第 j 个词向量 的余弦距离。

得到相关性矩阵 $\boldsymbol{M}_{n\times m}$ 后，首先，将检索语句和裁判文书中经预处理后的向量 $\boldsymbol{q}_i$ 与 $\boldsymbol{d}_j$ 经注意力机制计算，得到两者的注意力信号 $\boldsymbol{q}_i^{\text{att}}$ 与 $\boldsymbol{d}_j^{\text{att}}$；其次，对计算得到的注意向量与原向量作哈达玛积运算得到 $\boldsymbol{q}_i\circ\boldsymbol{q}_i^{\text{att}}$ 与 $\boldsymbol{d}_j\circ\boldsymbol{d}_j^{\text{att}}$；然后，分别将检索语句和裁判文书的原始向量、注意力向量和哈达玛积运算后的向量拼接得到修正后的向量 $\tilde{\boldsymbol{q}}_i$ 与 $\tilde{\boldsymbol{d}}_j$；最后，对修正后的向量 $\tilde{\boldsymbol{q}}_i$ 与 $\tilde{\boldsymbol{d}}_j$ 再次利用余弦进行相关性计算，可得到检索语句和裁判文书经过注意力修正后的相关性矩阵 $\tilde{\boldsymbol{M}}_{n\times m}$。

$$\boldsymbol{q}_i^{\text{att}}=\sum_{j=1}^{m}\frac{\mathrm{e}^{M_{i,j}}}{\sum_{l=1}^{m}\mathrm{e}^{M_{i,l}}}\boldsymbol{d}_j \tag{4}$$

$$\boldsymbol{d}_j^{\text{att}}=\sum_{i=1}^{n}\frac{\mathrm{e}^{M_{i,j}}}{\sum_{l=1}^{n}\mathrm{e}^{M_{l,j}}}\boldsymbol{q}_i \tag{5}$$

$$\boldsymbol{q}_i\circ\boldsymbol{q}_i^{\text{att}}=\left[\boldsymbol{q}_{i_1}\cdot\boldsymbol{q}_{i_1}^{\text{att}},\boldsymbol{q}_{i_2}\cdot\boldsymbol{q}_{i_2}^{\text{att}},\cdots,\boldsymbol{q}_{i_k}\cdot\boldsymbol{q}_{i_k}^{\text{att}}\right] \tag{6}$$

$$\boldsymbol{d}_j\circ\boldsymbol{d}_j^{\text{att}}=\left[\boldsymbol{d}_{j_1}\cdot\boldsymbol{d}_{j_1}^{\text{att}},\boldsymbol{d}_{j_2}\cdot\boldsymbol{d}_{j_2}^{\text{att}},\cdots,\boldsymbol{d}_{j_k}\cdot\boldsymbol{d}_{j_k}^{\text{att}}\right] \tag{7}$$

$$\tilde{\boldsymbol{q}}_i=\left[\boldsymbol{q}_i;\boldsymbol{q}_i^{\text{att}};\boldsymbol{q}_i\circ\boldsymbol{q}_i^{\text{att}}\right] \tag{8}$$

$$\tilde{\boldsymbol{d}}_i=\left[\boldsymbol{d}_j;\boldsymbol{d}_j^{\text{att}};\boldsymbol{d}_j\circ\boldsymbol{d}_j^{\text{att}}\right] \tag{9}$$

$$\tilde{M}_{i,j}=\cos\left(\tilde{\boldsymbol{q}}_i,\tilde{\boldsymbol{d}}_j\right) \tag{10}$$

其中，$\tilde{M}_{i,j}$ 为相关性矩阵 $\tilde{\boldsymbol{M}}_{n\times m}$ 内第 i 行第 j 列元素。

对相关性矩阵 $\tilde{\boldsymbol{M}}_{n\times m}$ 按行进行最大池化操作得到向量 $\boldsymbol{s}_{\text{word}}$，对向量 $\boldsymbol{s}_{\text{word}}$ 中每个元素 s_{word_i} 乘以该词的逆文档频率进行加权求和，得到检索语句和裁判文书词语级别的相关性得分 $score_{\text{word}}$。

$$s_{\text{word}_i}=Max\left(\tilde{\boldsymbol{M}}_{i,:}\right) \tag{11}$$

$$score_{\text{word}}=\sum_{i=1}^{n}idf_i^{\text{query}}\cdot s_{\text{word}_i} \tag{12}$$

$$idf_i^{\text{query}}=\log\frac{|D|}{df_i+1} \tag{13}$$

其中，idf_i^{query} 为词语 s_{word_i} 的逆文档频率，由数据库中的总文档数 $|D|$ 除以包含该关键字的文档数 df_i 并作对数运算得到。

3.2　词组级别匹配

法律领域文档涉及许多司法专家长期实践汇

总的专业术语和特定的句式表达，因此，不同的短语搭配往往表达不同的含义。这些短语容易被通用分词工具切割成独立的词语，割裂了短语之间的联系，造成原本语义信息的丢失，从而可能形成错误的表达。因此，为解决上述问题，词组级别匹配对词语级别匹配得到的相关性矩阵 $\boldsymbol{M}_{n\times m}$ 进行滑动窗口为 2×2 的平均池化操作，得到窗口内词语的关联信息，输出相关性矩阵 $\tilde{\boldsymbol{M}}^{\text{phrase}}_{(n-1)\times(m-1)}$，$\tilde{M}^{\text{phrase}}_{wi,wj}$ 为矩阵内第 i 行第 j 列元素。

$$\tilde{M}^{\text{phrase}}_{wi,wj}=\left(M^{\text{word}}_{wi,wj}+M^{\text{word}}_{wi+1,wj}+M^{\text{word}}_{wi,wj+1}+M^{\text{word}}_{wi+1,wj+1}\right)/4 \quad wi\in[1,n-1],wj\in[1,m-1] \tag{14}$$

对矩阵 $\tilde{\boldsymbol{M}}^{\text{phrase}}_{(n-1)\times(m-1)}$ 按行进行最大池化操作得到向量 $\boldsymbol{s}_{\text{phrase}}$，对向量 $\boldsymbol{s}_{\text{phrase}}$ 中每个元素 s_{phrase_i} 乘以该词的逆文档频率并加权求和，得到检索语句和文书词组级别的相关性得分 $score_{\text{phrase}}$。

$$s_{\text{phrase}_i}=Max\left(\tilde{\boldsymbol{M}}^{\text{phrase}}_{wi,:}\right) \tag{15}$$

$$score_{\text{phrase}}=\sum_{i=1}^{n}idf_i^{\text{query}}\cdot s_{\text{phrase}_i} \tag{16}$$

3.3 法律概念级别匹配

对律所提供的 8 000 万份法律文书经分词和去停用词处理后进行特征提取，得到案由数据共计 1 427 条，整合律所提供的 1 392 条专业案由知识数据，两份数据集进行去重整理后，得到刑事、民事、行政案由信息共计 1 541 条。结合官方法律案由知识信息和提取的案由关键词数据，构建法律案由知识词典。利用 TransE[28]算法将知识词典中的实体信息、实体间的关系映射到连续低维向量空间，并表示为三元组($\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}$)。其中，$\boldsymbol{h}, \boldsymbol{t}$ 分别表示实体向量信息；$\boldsymbol{r}$ 为关系向量，代表实体向量 $\boldsymbol{h}, \boldsymbol{t}$ 间的翻译。基于实体与实体间关系的向量表示，通过训练调整 $\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}$，最终使得 $\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$。其中，$\boldsymbol{t}$ 为 $\boldsymbol{h}+\boldsymbol{r}$ 的最近邻，且 $\boldsymbol{t}$ 与 $\boldsymbol{h}+\boldsymbol{r}$ 的距离足够远。图 2 为构建的部分法律知识词典示例。

提取预处理后检索语句和裁判文书中出现在知识词典中的实体数据，得到检索语句和裁判文书的实体向量表达矩阵，分别记为 $\boldsymbol{Q}^{\text{Ent}}_{n\times k}$，$\boldsymbol{D}^{\text{Ent}}_{m\times k}$。其中，$n$ 为检索语句在知识词典中的实体词语个数，序列中第 i 个实体词向量通过 TransE 算法输出为 $\boldsymbol{q}_i^{\text{Ent}}$，其维度为 k；m 为数据库中文档在知识词典中的实体词语个数，序列中第 j 个实体词向量通过 TransE 算法输出为 $\boldsymbol{d}_j^{\text{Ent}}$，其维度为 k。

$$\boldsymbol{Q}^{\text{Ent}}_{n\times k}=\left[\boldsymbol{q}_1^{\text{Ent}},\boldsymbol{q}_2^{\text{Ent}},\cdots,\boldsymbol{q}_{n-1}^{\text{Ent}},\boldsymbol{q}_n^{\text{Ent}}\right] \tag{17}$$

$$\boldsymbol{D}^{\text{Ent}}_{m\times k}=\left[\boldsymbol{d}_1^{\text{Ent}},\boldsymbol{d}_2^{\text{Ent}},\cdots,\boldsymbol{d}_{m-1}^{\text{Ent}},\boldsymbol{d}_m^{\text{Ent}}\right] \tag{18}$$

将 $\boldsymbol{Q}^{\text{Ent}}_{n\times k}$，$\boldsymbol{D}^{\text{Ent}}_{m\times k}$ 内元素两两相互计算余弦相似度得到相似度矩阵 $\boldsymbol{M}^{\text{Ent}}_{n\times m}$，其中第 i 行第 j 列元素记为

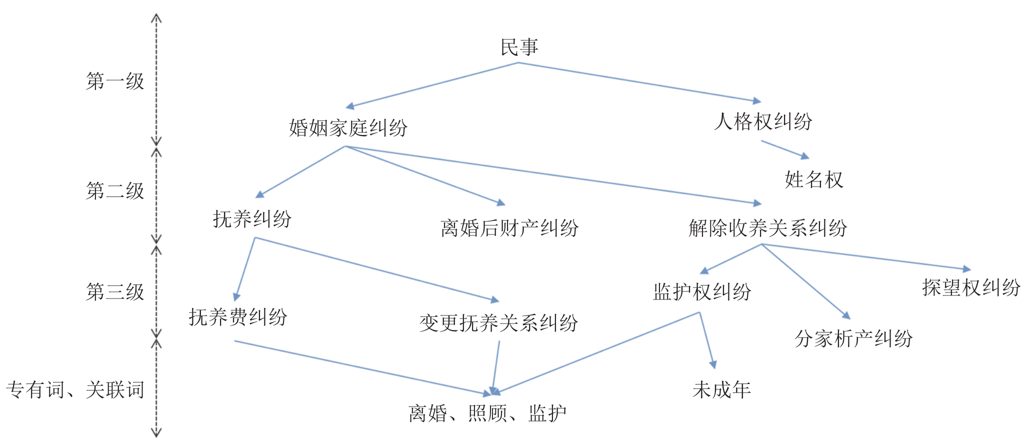


图 2 法律知识词典示例

Fig. 2 The example of legal knowledge base

$M_{i,j}^{\mathrm{Ent}}$。通过注意力机制得到检索语句实体和文档实体的注意力信号 $\boldsymbol{q}_i^{\mathrm{EntAtt}}$ 与 $\boldsymbol{d}_j^{\mathrm{EntAtt}}$，通过余弦相似度计算得到修正后的实体相关性矩阵 $\tilde{\boldsymbol{M}}_{n\times m}^{\mathrm{Ent}}$。对相关性矩阵进行行方向最大池化操作得到向量 $\boldsymbol{s}_{\mathrm{entity}}$，对向量 $\boldsymbol{s}_{\mathrm{entity}}$ 中每个元素 s_{entity_i} 乘以该词的逆文档频率进行加权求和，得到检索语句和裁判文书法律概念级别的相关性得分 $score_{\mathrm{entity}}$。

$$M_{i,j}^{\mathrm{Ent}}=\cos\left(\boldsymbol{q}_i^{\mathrm{Ent}},\boldsymbol{d}_j^{\mathrm{Ent}}\right) \tag{19}$$

$$\boldsymbol{q}_i^{\mathrm{EntAtt}}=\sum_{j=1}^{m}\frac{\mathrm{e}^{M_{i,j}^{\mathrm{Ent}}}}{\sum_{l=1}^{m}\mathrm{e}^{M_{i,l}^{\mathrm{Ent}}}}\boldsymbol{d}_j \tag{20}$$

$$\boldsymbol{d}_j^{\mathrm{EntAtt}}=\sum_{i=1}^{n}\frac{\mathrm{e}^{\boldsymbol{M}_{i,j}^{\mathrm{Ent}}}}{\sum_{l=1}^{n}\mathrm{e}^{\boldsymbol{M}_{l,j}^{\mathrm{Ent}}}}\boldsymbol{q}_i \tag{21}$$

$$\tilde{M}_{i,j}^{\mathrm{Ent}}=\cos\left(\boldsymbol{q}_i^{\mathrm{EntAtt}},\boldsymbol{d}_j^{\mathrm{EntAtt}}\right) \tag{22}$$

$$s_{\mathrm{entity}_i}=Max\left(\tilde{\boldsymbol{M}}_{i,:}\right) \tag{23}$$

$$score_{\mathrm{entity}}=\sum_{i=1}^{n}idf_i^{\mathrm{query}}\cdot s_{\mathrm{entity}_i} \tag{24}$$

将词语级别匹配得分 $score_{\mathrm{word}}$，词组级别匹配得分 $score_{\mathrm{phrase}}$，法律概念级别匹配得分 $score_{\mathrm{entity}}$ 按照 α,β,γ 的比例进行加权求和，得到检索语句与文档的最终相关性得分 $score_{\mathrm{rank}}$。

$$score_{\mathrm{rank}}=\alpha\cdot score_{\mathrm{word}}+\beta\cdot score_{\mathrm{phrase}}+\gamma\cdot score_{\mathrm{entity}} \tag{25}$$

$$\alpha,\beta,\gamma\in[0,1],\alpha+\beta+\gamma=1$$

4　结果分析与评估

4.1　数据集与评价指标

实验检索语句数据集由法律行业专业从业人员进行整理，其利用专业知识和长期实践经验分别从债权债务、劳动纠纷、基础设施、婚姻家事、房地产、投资并购和知识产权 7 个方面提炼检索语句共计 958 条；将检索语句输入 Elasticsearch 法律文书数据库，数据库包含法律文书 8 000 万份，每条语句召回的前 30 份相关文档交由法律行业专业从业人员进行相关性标记，共计标记文书 23 196 份，构成检索文档数据集。

数据集描述如表 1 所示，将 958 条检索语句按 8∶2 的比例拆分成训练集和测试集，其中，训练集共计包含文档 18 589 份，平均每条检索语句标记正相关文档 10.64 份，负相关文档 13.78 份，每份文档预处理前标题平均长度为 32.09 字，本院认为部分平均长度 1 188.99 字；测试集共计包含文档 4 648 份，平均每条检索语句标记正相关文档 11.07 份，负相关文档 13.19 份，每份文档预处理前标题平均长度为 30.73 字，本院认为部分平均长度 1 209.19 字。

表 1　法律裁判文书检索数据集描述

Table 1　Statistics on legal case retrieval dataset

	法律裁判文书数据集	
	训练集	测试集
语言	中文	中文
检索语句	766	192
文档数	18 589	4 648
每条语句标记文档数	＋10.64,－13.78	＋11.07,－13.19
标题平均字数	32.09	30.73
本院认为平均字数	1 188.99	1 209.19

实验采用的评价指标为检索模型通用评价指标分别为 MAP, MRR, NDCG@10。

平均精度均值(Mean Average Precision，MAP)，从多个查询上反映模型的整体检索性能。Q 为相关检索的次数，N 为每次检索的相关文档数，$position$ 为相关文档 i 的排名位置，相关文档排名越靠前则 $AvgP$ 越高，MAP 值也就越高。若系统没有返回相关文档，那么 $AvgP$ 值为 0，具体计算公式如下：

$$MAP=\frac{1}{|Q|}\sum_{q=1}^{|Q|}AvgP\left(q\right) \tag{26}$$

$$AvgP = \frac{1}{N}\sum_{i=1}^{N}\frac{i}{position(i)} \tag{27}$$

平均倒数排名(Mean Reciprocal Rank，MRR)，根据查询排序的正确性，对查询的结果进行评价。对每个查询 Q_i，记第一个相关结果的位置为 k_i，则其得分记为 $\frac{1}{rank_i}$，对所有的查询分别计算得分并取均值，具体计算公式如下：

$$MRR = \frac{1}{|Q|}\sum_{q=1}^{|Q|}\frac{1}{rank(q)} \tag{28}$$

归一化折损累计增益(Normalized Discounted Cumulate Gain，NDCG)，对面向打分的个性化推荐系统或相关检索程序的有效性进行度量。指标考虑检索结果的位置因素，排名越靠前的结果价值越高。其中，$|REL|$ 表示结果按照相关性从大到小排序，即按照最优的方式对结果进行排序后取前 P 个结果组成的集合；rel_i 表示第 i 个结果的得分。最终指标结果在 0～1 之间，越接近 1 说明检索越准确。具体计算公式如下：

$$NDCG = \frac{\sum_{i=1}^{p}\frac{rel_i}{\log_2(i+1)}}{\sum_{i=1}^{|REL|}\frac{rel_i}{\log_2(i+1)}} \tag{29}$$

4.2 实验结果

本文利用表 1 的法律数据集进行法律文书检索实验，数据集中检索语句与法律文书的相关性均通过人工标注，以保证检索结果的正确性。为证明实验结果的有效性，本文还选用相同的法律数据集，将 ULRM 模型与过去发表的无监督检索模型和在法律领域预微调的训练模型进行对比实验。对比的基准模型有：

(1) Jones[10]提出的 TF-IDF 模型，将检索文档和检索语句转换为特征向量，计算词频和逆文档频率，用以评估检索关键词对语料库中文件的重要程度，还利用加权技术得到检索语句与检索文档的相关性得分。

(2) Roberston[11]提出的 BM25 模型，在 TF-IDF 模型的基础上增加了可调节的参数，控制最终得分对单一词频的敏感程度，计算检索语句中每个词与每个文档的相关性得分，然后加权求和，同时引入文档长度和平均文档长度之比作为惩罚因子，最终得到相关性得分。

(3) 查询似然模型(Query Likelihood Model，QLM)[28]为集合中每个文档构建对应的语言模型，利用贝叶斯算法计算给定的查询语句在文档集合中的生成概率，然后得到似然排序。

(4) Gysel[29]提出的 NVSM 模型以无监督的的方式使用梯度下降算法，从零开始学习检索语句和文档的低维向量表示，最后根据文档和检索语句向量间的相似性对文档进行排序。

(5) Shao 等[30]提出的 BERT-PLI 模型，利用 BERT 捕获段落级别的语义关系，然后通过汇总段落级别的交互来推断两个案例之间的相关性。本文使用划分的法律检索数据集对 BERT 模型进行微调，使其适用于法律场景。

4.3 讨论与分析

表 2 为已有基线模型与 ULRM 模型在法律裁判文书数据集上的检索实验结果。TF-IDF 与 BM25 模型指标相近，QLM 模型指标则较两者表现稍差，由此可得，在法律裁判文书检索领域，利用文档与检索语句之间的相关性建模方式效果略差于基于关键词匹配的 TF-IDF 与 BM25 模型。与已有无监督的基线模型相比，ULRM 模型利用外部法律知识提高了法律知识的匹配能力，综合词语级别、词组级别和法律概念级别 3 个方面的匹配信号，在 MAP、MRR 和 NDCG@10 指标上有明显提升；与基于 BERT 并在法律数据集上微调的预训练模型 BERT-PLI 相比，ULRM 模型在 MRR 与 NDCG@10 指标上仍有小幅优势。本实验结果证明，ULRM 模型在法律文书检索数据集上检索的有效性。

表 2　法律裁判文书数据集检索实验结果

Table 2　Retrieval performance on legal case retrieval dataset

模型	MAP	MRR	NDCG@10
TF-IDF	0.486 3	0.563 6	0.486 8
BM25	0.484 4	0.552 6	0.481 7
QLM	0.481 1	0.539 1	0.479 0
NVSM	0.506 2	0.601 8	0.530 7
BERT-PLI	0.671 7	0.753 9	0.675 5
ULRM	0.666 1	0.754 5	0.679 4

ULRM 模型最终得分由词语级别匹配、词组级别匹配和法律概念级别匹配 3 个方面的评分综合得出。为了评估各模块的重要性，本文还对 ULRM 模型进行了消融实验，即在单次实验中去除一个模块后观测实验指标的变化情况，实验结果如图 3 和表 3 所示。从表 3 中可以发现：

表 3　法律数据集检索消融实验结果

Table 3　Ablation experiments on legal case retrieval dataset

模型	MAP	MRR	NDCG@10
$\mathrm{ULRM}_{\mathrm{phrase\_concept}}$	0.656 6	0.739 5	0.668 1
$\mathrm{ULRM}_{\mathrm{word\_phrase}}$	0.659 7	0.747 8	0.668 3
$\mathrm{ULRM}_{\mathrm{word\_concept}}$	0.663 8	0.750 2	0.676 5
$\mathrm{ULRM}_{\mathrm{full}}$	0.666 1	0.754 5	0.679 4

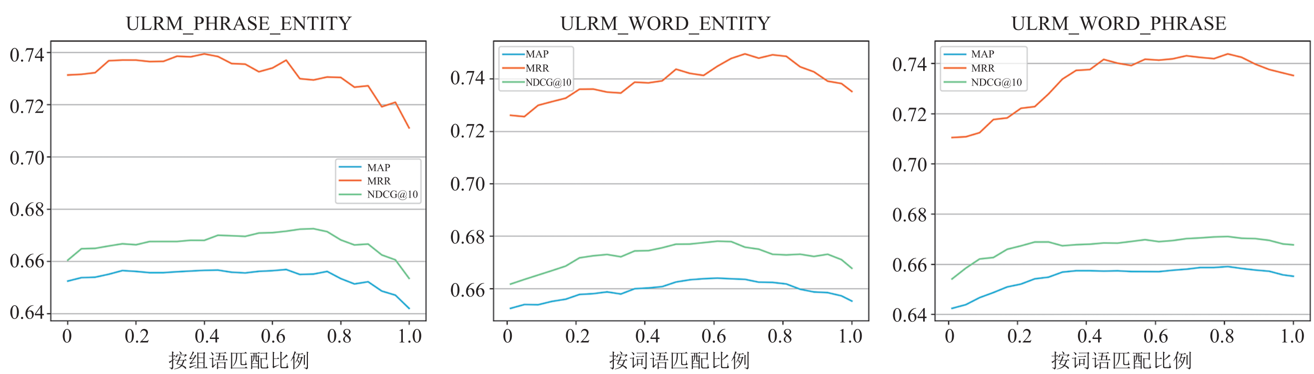
(1) 仅使用一个模块时，按词语级别匹配 3 个指标分别为 MAP：0.655 3、MRR：0.735 2、NDCG@10：0.667 8，较其他模块下降最少。按词组级别匹配 3 个指标分别为 MAP：0.642 1、MRR：0.711 2、NDCG@10：0.653 6，较其他模块下降最多。按法律概念匹配 3 个指标分别为 MAP：0.652 3、MRR：0.731 4、NDCG@10：0.660 5。综上所述，可知按词语级别匹配对 ULRM 模型的作用最大。

(2) 当禁用按词语级别匹配时，随着按词组级别匹配比例的增加，MRR 指标趋于平缓，当比例超过 0.4 时逐渐降低；MAP 和 NDCG@10 指标随比例的增加呈缓慢上升趋势，当比例在 0.7 时趋近最大值，然后快速下降。

(3) 当禁用按词组级别匹配时，随着按词语级别匹配比例的增加，3 个指标变化趋势基本一致，呈逐渐上升趋势，在 0.7 附近时，3 个指标均逼近最大值后逐渐下降，其中，MRR 指标下降明显。

(4) 当禁用按法律概念级别匹配时，随着按词语级别匹配比例的增加，MRR 指标变化明显，呈快速上升趋势，在 0.8 附近接近最大值后缓慢下降。MAP、NDCG@10 指标先呈缓慢上升趋势，当比例超过 0.2 时趋于平缓。

本实验验证了按词语级别匹配、按词组级别匹配和按法律概念级别匹配对于模型的指标均有积极作用。当使用单一模块时模型 3 个指标的表现普遍降低 0.1～0.2，引入法律知识词



(a) 按词组和法律概念级别匹配　(b) 按词语和法律概念级别匹配　(c) 按词语和词组级别匹配

图 3　ULRM 模型消融实验结果

Fig. 3　ULRM ablation experiment

表 4 法律裁判文书检索示例

Table 4 An example of legal case retrieval

模型	文书标题	文书案由	是否相关
TF-IDF	保康县国土资源局与扬州**公司建设用地使用权出让合同纠纷一审民事判决书	建设用地使用权出让合同纠纷	是
BM25	丰宁**公司与丰宁**国土资源局行政处理决定二审行政判决书	行政复议	是
QLM	保康县国土资源局与扬州**公司建设用地使用权出让合同纠纷一审民事判决书	建设用地使用权出让合同纠纷	是
NVSM	阳泉**公司与阳泉国土资源局、阳泉市人民政府行政裁决二审行政判决书	行政裁决	是
BERT-PLI	董*、朱*等与尹**建设用地使用权转让合同纠纷一审民事判决书	建设用地使用权转让合同纠纷	是
ULRM	汉江集团**公司与湖北**城乡**公司房地产开发经营合同纠纷一审民事判决书	房地产开发经营合同纠纷	是

典中的实体信息用更为准确的语义描述进行匹配，可以将按词语级别匹配的 MAP、MRR、NDCG@10 分别从 0.655 3、0.735 2、0.667 9 提升至 0.663 8、0.750 2、0.676 5；将按词组级别匹配的 MAP、MRR、NDCG@10 分别从 0.642 1、0.711 2、0.653 6 提升至 0.656 6、0.739 5、0.668 1。控制一个模块后逐渐变化各模块的比例时模型表现呈献先上升后下降趋势。对比各模块占比对实验指标的影响，选用按词语级别匹配比例 0.42，按词组级别匹配比例 0.39，按法律概念级别匹配比例 0.19 时可以获得接近最优的指标，此时 MAP 为 0.666 1，MRR 为 0.754 5，NDCG@10 为 0.679 4。表 4 为法律裁判文书检索示例，其以“土地出让合同约定了建设规模、容积率，但建成物业的建设规划、容积率超出土地出让合同约定的法律后果？”为检索问题，各模型经检索排序后的第一个结果标题及案由信息如表 4 所示。

5 结 论

在法律文书检索领域，由于文书内容的特殊性，往往缺乏大量的标注数据，从而难以通过训练得到效果优异的深度学习模型。此外，标注训练数据需要耗费大量法律行业从业人员的时间及精力，人工成本较高。因此，本文提出了基于无监督的法律裁判文书检索模型 ULRM，综合词语级别、词组级别和法律概念级别 3 个方面的打分结果，来对召回文书进行重排名：在词语级别，利用注意力机制获取用户输入信息和法律文书的关键词信号；在词组级别，引入平均池化层模拟短语匹配从而获取更多关键词匹配信号；在法律概念级别，考虑法律专业术语实体信息从而更准确地获取匹配信号，提升模型的检索效果。

模型使用无监督的方式避免了对大量标注数据的依赖，并通过实验验证，该模型在法律裁判文书数据集上的 MAP、MRR、NDCG@10 指标远超已有基线模型，与基于 BERT 并在法律数据集上微调的预训练模型 BERT-PLI 结果相近，能够对召回文书进行有效重排序。由于模型 3 个模块主要使用关键字信息进行检索，对于语料库中缺少的内容无法识别，同时由于词袋模型的使用无法关联上下文语义，对模型指标的提升仍有局限。因此，将模型部署上线，利用无监督模型获取用户的检索及点击数据，对于标注数据的获取及后期进行有监督模型的训练具有重要意义。

参 考 文 献

[1] Rabelo J, Kim MY, Goebel R, et al. A summary of the COLIEE 2019 competition [J]. New Frontiers

in Artificial Intelligence, 2019, 12331: 34-49.

[2] Harb HM, Fouad KM, Nagdy NM. Semantic retrieval approach for web documents [J]. International Journal of Advanced Computer Science and Applications, 2011, 2(9): 67-76.

[3] ManningCD, Raghavan P, Schütze H. Introduction to information retrieval [M]. Cambridge: Cambridge University Press, 2008.

[4] Salton G, Fox EA, Wu H. Extended Boolean information retrieval [J]. Communications of the ACM, 1983, 26(11): 1022-1036.

[5] Salton G,WongA,Yang CS. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11): 613-620.

[6] Robertson SE, Jones KS. Relevance weighting of search terms [J]. Journal of the Association for Information Science, 1976, 27(3): 129-146.

[7] Rosenfeld R. Two decades of statistical language modeling: where do we go from here? [J]. Proceedings of the IEEE, 2000, 88(8): 1270-1278.

[8] Liu TY. Learning to rank for information retrieval [J]. Foundations and Trends in Information Retrieval, 2009, 3(3): 225-331.

[9] Guo JF, Fan YX, Pang L, et al. A deep look into neural ranking models for information retrieval [J]. Information Processing and Management, 2020, 57(6): 102067.

[10] Jones KS. A statistical interpretation of term specificity and its application in retrieval [J]. Journal of Documentation, 1972, 28(1): 11-21.

[11] Robertson SE, Walker S, Jones S, et al. Okapi at TREC-3 [C] // Proceedings of the Overview of the Third Text REtrieval Conference, 1994: 109-123.

[12] Huang PS, He XD, Gao JF, et al. Learning deep structured semantic models for Web search using clickthrough data [C] // Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, 2013: 2333-2338.

[13] Shen YL, He XD, Gao JF, et al. A latent semantic model with convolutional-pooling structure for information retrieval [C] // Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 2014: 101-110.

[14] Palangi H, Deng L, Shen YL, et al. Deep sentence embedding using Long Short-Term Memory networks: analysis and application to information retrieval [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(4): 694-707.

[15] Hu BT, Lu ZD, Li H, et al. Convolutional neural network architectures for matching natural language sentences [C] // Advances in Neural Information Processing Systems, 2014: 2042-2050.

[16] Guo JF, Fan YX, Ai QY, et al. A deep relevance matching model for ad-hoc retrieval [C] // Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, 2016: 55-64.

[17] Opijnen MV, Santos C. On the concept of relevance in legal information retrieval [J]. Artificial Intelligence and Law, 2017, 25(1): 65-87.

[18] Zhong HX, Xiao CJ, Tu CC, et al. How does NLP benefit legal system: a summary of Legal Artificial Intelligence [C] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 5218-5230.

[19] Ashley KD. Artificial intelligence and legal analytics: new tools for law practice in the digital age [M]. Cambridge: Cambridge University Press, 2017.

[20] Sugathadasa K, Ayesha B, Silva N, et al. Legal document retrieval using document vector embeddings and deep learning [C] // Proceedings of the Science and Information Conference, 2018: 160-175.

[21] Chalkidis I, Kampas D. Deep learning in law: early adaptation and legal word embeddings trained on large corpora [J]. Artificial Intelligence and Law, 2019, 27(2): 171-198.

[22] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C] // Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013: 3111-3119.

[23] Tran V, Nguyen ML, Satoh K. Building legal

case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model [C] // Proceedings of the 17th International Conference on Artificial Intelligence and Law, 2019: 275-282.

[24] Liu ZH, Xiong CY, Sun MS, et al. Entity-Duet Neural Ranking: understanding the role of knowledge graph semantics in neural information retrieval [Z/OL]. arXiv Preprint, arXiv: 1805.07591, 2018.

[25] Zhang Y, Wang D, Zhang Y. Neural IR meets graph embedding: a ranking model for product search [C] // Proceedings of the World Wide Web Conference, 2019: 2390-2400.

[26] Duan W, Li L. An external knowledge enhanced multi-label charge prediction approach with label number learning [Z/OL]. arXiv Preprint, arXiv: 1907.02205, 2019.

[27] Bordes A, Usunier N, Duran AG, et al. Translating embeddings for modeling multi-relational data [C] // Proceedings of the Annual Conference on Neural Information Processing Systems, 2013: 2799-2807.

[28] Ogawa K, Murahashi T, Taguchi H, et al. Spoken document retrieval using neighboring documents and extended language models for query likelihood model [C] // Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, 2016: 186-190.

[29] Gysel CV, Rijke MD, Evangelos K, et al. Neural Vector Spaces for unsupervised information retrieval [J]. ACM Transactions on Information Systems, 2018, 36(4): 38.

[30] Shao YQ, Mao JX, Liu YQ, et al. BERT-PLI: modeling paragraph-level interactions for legal case retrieval [C] // Proceedings of the 29th International Joint Conference on Artificial Intelligence, 2020: 3501-3507.