

# GSGD:一种基于 BERT 与本体推理的自动分级系统

王珊珊<sup>1,2</sup>, 邹佳<sup>1,2</sup>, 程序<sup>1,2</sup>, 刘汪洋<sup>1,2</sup>, 蔡惠民<sup>1,2</sup>

(1. 中电科大数据研究院有限公司, 贵州 贵阳 550022;  
2. 提升政府治理能力大数据应用技术国家工程实验室, 贵州 贵阳 550022)

**摘要:**政府数据资源分级管理是政府数据共享开放和数据治理的关键性工作。由于数据资源规模大,分级体系不完善,工具缺乏,使得该工作多由人工进行,导致支撑依据不足、主观性强、精确性差、成效不足。文中设计并实现了基于政策法规、典型案例的政府数据自动分级系统—GSGD (grading system for government data)。首先,利用政策法规以及典型案例构建本体库,根据分级目标以及构建的本体特性,构建自定义推理规则;再通过 BERT 获得输入数据与关键词的语义特征词/句向量,并计算向量之间的余弦相似度;最后对相似度较高的关键词,采用 Jena 对政策法规库以及典型案例库进行查询推理得到分级结果以及分级依据,实现对政府数据的自动化分级,提高分级工作效率。通过实验对比分析,验证了该方法的有效性。

**关键词:**数据分级;政府数据;BERT;法律本体;余弦相似度

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2020)08-0097-06

doi:10.3969/j.issn.1673-629X.2020.08.016

## An Automatic Grading System Based on BERT and Ontology Reasoning

WANG Shan-shan<sup>1,2</sup>, ZOU Jia<sup>1,2</sup>, CHENG Xu<sup>1,2</sup>, LIU Wang-yang<sup>1,2</sup>, CAI Hui-min<sup>1,2</sup>

(1. CETC Big Data Research Institute Co., Ltd., Guiyang 550022, China;

2. National Engineering Laboratory for Big Data Application in Improving Government Governance Capabilities, Guiyang 550022, China)

**Abstract:** Grading of government data resources is the key work of government data sharing and opening. Due to the large scale of data resources, imperfect classification system and lack of tools, this work is mostly carried out manually, which leads to insufficient supporting basis, strong subjectivity, poor accuracy and insufficient effectiveness. Therefore, we design and implement GSGD, an automatic grading system for government data based on policies, regulations and typical cases. Firstly, policies and regulations as well as typical cases are used to build ontology, and custom inference rules are built according to grading work and the ontology characteristics. Then, the semantic features word/sentence vectors of the input data and keywords are obtained through BERT, and cosine similarity between the vectors is calculated. Finally, for keywords with high similarity, Jena is used to query and reason the policy and regulation database and typical case database to obtain grading results and basis, which helps automatically to grade the data. The effectiveness of the method is verified by experiment.

**Key words:** data grading; government data; BERT; legal ontology; cosine similarity

## 0 引言

政府数据分级管理,能够明确政府数据的范围边界和使用方式,是政府数据治理的关键性工作,为数据共享开放提供依据<sup>[1-2]</sup>。国务院2015年9月5日印发的《促进大数据发展行动纲要》(国发〔2015〕50号)的主要任务中明确提出要大力推动政府部门数据共享,稳步推动公共数据资源开放。国务院办公厅于2017

年5月18日印发并实施《政务信息系统整合共享实施方案》(国办发〔2017〕39号),提出了加快推进政务信息系统整合共享。2018年1月12日,贵阳市发布《贵阳市政府数据共享开放实施办法》,用以协调解决政府数据共享开放有关重大问题。

根据《政务信息资源共享管理暂行办法》、《贵州省政务数据资源管理暂行办法》、《贵阳市政府数据共

收稿日期:2019-10-10

修回日期:2020-02-13

基金项目:天津市新一代人工智能科技重大专项(18ZXZNGX00370)

作者简介:王珊珊(1993-),女,硕士,研究方向为语义 Web、机器学习。

享开放实施办法》,政府数据分级主要是对数据在开放和共享两个方向进行分级;共享级别分别为无条件共享、有条件共享、不予共享三大等级,开放级别分别为无条件开放、依申请开放和不予开放三大等级。

目前,政府数据分级工作多为人工操作,然而,随着政府数据的增长,人工标注已不能满足分级工作要求,带来了很多问题。由于分级政策法规条款较多,人工对大量的数据进行分级时需不停查阅相关规定导致工作量大、效率低;同时人为理解政策法规具有较强的主观性,导致现有人工分级工作精确性差、较为主观等。由于分级工作涉及领域较广,例如:安全生产、健康保障、信用体系等,且需要政策法规依据支撑结果,因此传统的分类方法不足以支撑分级工作。

法律本体能够对法律法规进行条理的梳理、描述;还可通过自定义规则,以满足个性化推理需求。Valente 从法律的社会角色和功能出发,提出了 FOLaw (functional ontology for law)<sup>[3]</sup> 法律本体。Breuker<sup>[4]</sup> 创建了 LRI-Core 法律本体模型。汤庸等结合了许多研究,提出了新的本体模型 DOLegal<sup>[5]</sup>。贾君枝<sup>[6]</sup> 等以专业人员参与为核心,提出了一种新的法律框架网络知识本体模型。卢明纯<sup>[7]</sup> 在结合国内外研究成果的基础上,提出了一种新的本体模型,并设计了原型系统。余贵清等<sup>[8]</sup> 基于历史案例本体知识库构建了刑事审判案例推理模型。姜赢等<sup>[9]</sup> 构建了医疗卫生政策法律知识库,以方便对政策法律进行管理。Thammaboosadee 等<sup>[10]</sup> 根据泰国刑法典提出了一个判决系统。上述研究大多针对《刑法》等法律且推理规则多关注于行为处罚措施,涉及法律内容较为单一。

本体的语义匹配技术较多,有基于模式的匹配、基于概念图的匹配,以概念分类为基础的学习策略等;贾君枝等在充分考虑法律语言的模糊性上,结合了相关技术,提出了基于法律框架网络本体的语义匹配的基本思路;但基于框架网络的语义匹配更适合应用于范围界限较为清晰的领域<sup>[11]</sup>。

随着大数据等技术的发展,采用大数据、人工智能等方法对政府数据自动进行分级已成必然趋势。因此,文中以《中华人民共和国政府信息公开条例》、《政务信息资源共享管理暂行办法》以及贵州省、贵阳市地方法规、标准等作为政策法规依据,以某些省市开放平台中的典型案例作为案例数据,设计并实现了政府数据自动分级系统—GSGD,以解决现有人工分级支撑依据不足、主观性强、精确性差的问题。

## 1 系统框架

GSGD 由输入数据、基础能力、算法模型、结果输出四个部分构成,系统框架如图 1 所示。分级输入数

据格式为 xx 市政府各委办局“行政区 委办局名称 系统名称 表名称 字段名称”目录,输入数据样例见表 1。

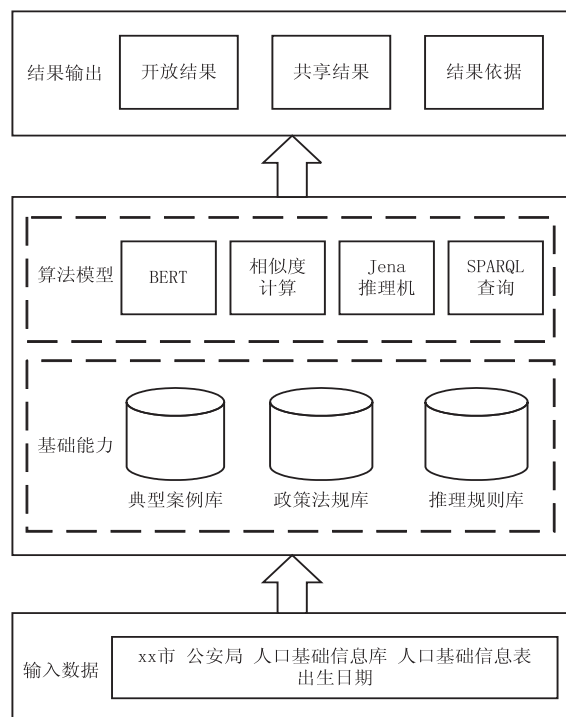


图 1 分级系统框架

表 1 输入数据样例与分级结果

输入数据	开放级别	共享级别
xx 市 公安局 人口基础信息库 人口基础信息表 出生日期	依申请开放	无条件共享
xx 市 卫计委 免疫规划信息管理系统 新生儿接种登记 接种日期	依申请开放	有条件共享

基础能力以及算法模型板块完成了数据中间处理过程。基础能力板块主要是政策法规库、典型案例库、推理规则库,文中分级结果以《中华人民共和国政府信息公开条例》、《政务信息资源共享管理暂行办法》以及《贵州省政务数据资源管理暂行办法》、《贵阳市政府数据共享开放条例》、《贵阳市政府数据共享开放实施办法》等贵州省、贵阳市地方法规、标准作为依据,构建政策法规库;以某些省市开放平台中的典型案例作为依据,构建典型案例库;根据政策法规库以及典型案例库中本体概念以及框架,设计自定义推理规则构成推理规则库。将政策法规库以及典型案例库中的关键词(例如:人事任免、健康保障等)提出作为分级关键词。算法模型板块由 BERT<sup>[12]</sup> 模型、相似度计算、Jena 推理机<sup>[13]</sup> 以及 SPARQL 查询<sup>[14]</sup> 构成;BERT 与相似度计算完成输入数据到政策法规库/典型案例库中关键词的映射过程;Jena 推理机以及 SPARQL 查询完成政策法规库/典型案例库中关键词到分级结果的推理分析过程。结果输出模块将对算法模型模块的结果进行整理,并格式化输出,输出内容包括:开放结果、共

享结果以及结果依据。系统整体流程如图 2 所示。

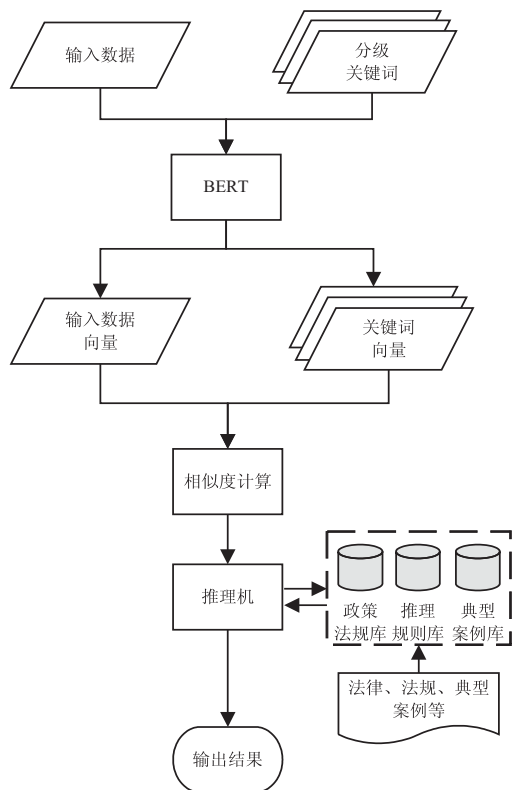


图 2 系统整体流程

## 2 系统模块设计

### 2.1 本体构建

文中采用 Protégé 作为构建本体工具,Protégé 是由斯坦福大学开发的本体编辑器,具有众多的插件。Protégé 能够直观地以树形层次目录结构显示本体,且操作简便,是目前使用最广泛的本体编辑器之一<sup>[15-16]</sup>。

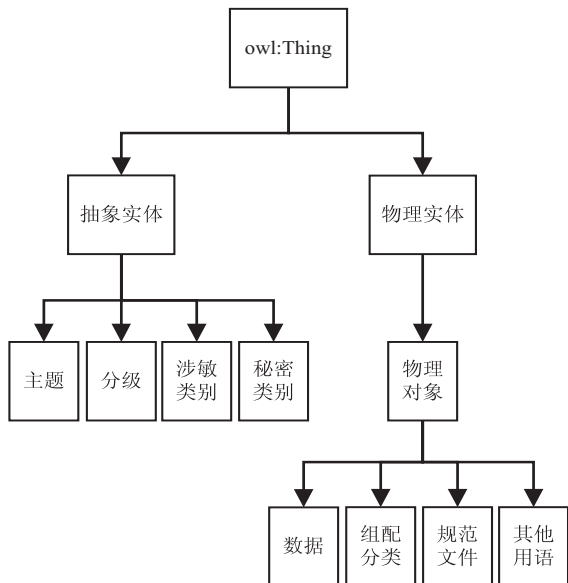


图 3 政策法规库本体框架

文中参考许多已有的研究,并结合分级工作的特

性,构建了分级政策法规库以及典型案例库。政策法规本体库顶层划分为两大概念:抽象实体和物理实体。抽象实体的子类有主题、分级、涉敏类别以及秘密类别,物理实体的子类有物理对象。根据贵阳市政府信息公开目录对政策法规进行概念提取,例如,组配分类中的子类有:人事信息、总结公报、规划计划等。规范文件可分为:宪法、法律、行政法规、地方性法规、部门规章、其他规范性文件,规范文件子类中各概念之间的效力级别采用“效力高于”这一对象属性进行描述<sup>[17]</sup>,详细的分类如图 3 所示。典型案例库采用与构建政策法规库相似的方式进行构建,典型案例库的本体框架如图 4 所示。构建数据为某些省市政府开放数据平台上获得的典型案例,例如:机动车驾驶证满分名单等。

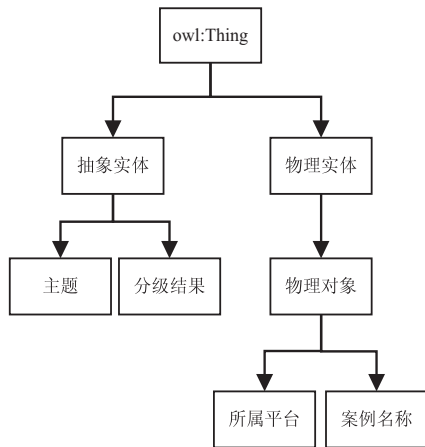


图 4 典型案例库本体框架

### 2.2 基于 BERT 的相似度计算

BERT (bidirectional encoder representations from transformers) 是基于深度双向 Transformer 的预训练模型,BERT 在训练任务中关注词前后的信息,生成融合了上下文信息的语义向量,因此,BERT 可以用于问答系统、命名实体识别、文本挖掘等任务中<sup>[12,18-20]</sup>。文中利用 BERT 获得精准的语义向量,并将语义向量用于输入数据以及分级关键词的相似度计算中。

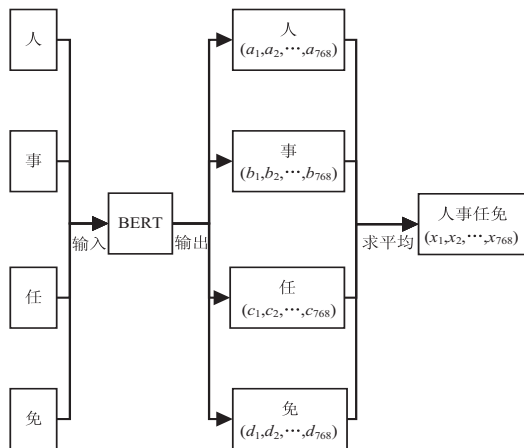


图 5 求词/句向量流程

通过计算输入数据中委办局名称、系统名称、表名

称、字段名称部分分别与分级关键词的词/句向量相似度,选取输入数据每个部分所对应相似度较高的关键词作为查询推理的输入。词/句向量采用 BERT 进行计算,将 BERT 模型的输出,即模型最后一层的输出,作为输入数据/关键词中每个字的字向量;对输入数据/关键词的字向量求平均,得到输入数据/关键词的词/句向量,流程如图 5 所示。

计算输入数据各部分的词/句向量与每个分级关键词的词/句向量的余弦相似度,并取输入数据各部分对应相似度最大的前两个关键词组成的关键词集合作为查询推理的输入。余弦相似度用两个向量夹角的余弦值作为衡量两个个体间差异的大小,更加注重两个向量在方向上的差异,较多地应用于文本相似度计算<sup>[21-22]</sup>;假设文档  $x = \langle x_1, x_2, \dots, x_n \rangle$ ,  $y = \langle y_1, y_2, \dots, y_n \rangle$ , 其余弦相似度为<sup>[23]</sup>:

$$\text{Sim}(x, y) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \cdot \|\vec{y}\|} \quad (1)$$

### 2.3 推理规则

文中使用 Jena 推理机完成本体查询以及推理模

块。Jena 是由 HP Labs 开发的 Java 开发,是一种开源的产生式规则的前向推理系统,可通过自定义规则完成个性化推理,通过 Jena 提供的 OWL API 接口、SPARQL 查询接口和本体推理机接口,可以实现基于本体智能应用程序<sup>[13,24-25]</sup>。

文中通过自定义的推理规则对通用规则进行扩展,满足对实际应用的个性化需求,本体中有间接关系的概念可通过规则的制订,经过推理最终被查询到。Jena 的推理规则分为前向规则和后向规则,文中使用的是前向规则,规则分为前提和结论,形式如下,其中 term 和 hterm 是三元组或扩展三元组<sup>[26-27]</sup>。

$$\text{term}, \dots, \text{term} \rightarrow \text{hterm}, \dots, \text{hterm} \quad (2)$$

表 2 列出了部分推理规则及其功能。由于一些政策法规条款内容较为相似,例如,贵阳市政府数据共享开放实施办法第二十五条与贵州省政务数据资源管理暂行办法第二十八条。因此文中采用规则对条款之间的关系进行处理,使得某一条款“继承”与其内容相似条款的关系,减轻人工构建本体时的工作量。虽然,文中所涉及的政策法规没有冲突,为防止随着政策法规

表 2 部分推理规则及其功能

功能	规则
内容相似 条款处理	<pre>[rule8:(? a rdfs:subClassOf ? b)(? b owl:onProperty law:内容相似)(? b owl:someValuesFrom ? c) (? c rdfs:subClassOf ? d)(? d owl:onProperty law:反应),(? d owl:someValuesFrom ? e) (? f rdf:type owl:Restriction)(? f owl:onProperty law:反应)(? f owl:someValuesFrom ? e)-&gt; (? a rdfs:subClassOf ? f)]</pre> <pre>[rule11:(? a rdfs:subClassOf ? b)(? b owl:onProperty law:内容相似)(? b owl:someValuesFrom ? c) (? c rdfs:subClassOf ? d)(? d owl:onProperty law:内容为)(? d owl:someValuesFrom ? e) (? f rdf:type owl:Restriction)(? f owl:onProperty law:内容为)(? f owl:someValuesFrom ? e)-&gt; (? a rdfs:subClassOf ? f)]</pre>
冲突检测	<pre>[rule14:(? law rdfs:subClassOf ? a)(? a owl:onProperty law:反应)(? a owl:someValuesFrom ? topic)(? law rdfs:subClassOf ? b)(? b owl:onProperty law:内容为)(? b owl:someValuesFrom ? result)(? law2 rdfs: subClassOf ? c)(? c owl:onProperty law:反应)(? c owl:someValuesFrom ? topic) (? law2 rdfs:subClassOf ? d)(? d owl:onProperty law:内容为)(? d owl:someValuesFrom ? result2)(? result owl:disjointWith ? result2)(? law rdfs:subClassOf ? e)(? e owl:onProperty law:效力高于)(? e owl: allValuesFrom ? law2)(? f rdf:type owl:Restriction)(? f owl:onProperty law:应) (? f owl:allValuesFrom ? result)-&gt;(? topic rdfs:subClassOf ? f)]</pre> <pre>[rule15:(? law rdfs:subClassOf ? a)(? a owl:onProperty law:反应)(? a owl:someValuesFrom ? topic)(? law rdfs:subClassOf ? b)(? b owl:onProperty law:内容为)(? b owl:someValuesFrom ? result)(? law2 rdfs:subClassOf ? c)(? c owl:onProperty law:反应)(? c owl:someValuesFrom ? topic) (? law2 rdfs:subClassOf ? d)(? d owl:onProperty law:内容为)(? d owl:someValuesFrom ? result2)(? result owl:disjointWith ? result2)-&gt;(? law owl:disjointWith ? law2)]</pre>
政策法规库 获得分级结果	<pre>[rule12:(? a rdfs:subClassOf ? b)(? b owl:onProperty law:内容为)(? b owl:someValuesFrom ? c) (? a rdfs:subClassOf ? d)(? d owl:onProperty law:反应)(? d owl:someValuesFrom ? e) (? f rdf:type owl:Restriction)(? f owl:onProperty law:应)(? f owl:allValuesFrom ? c)-&gt; (? e rdfs:subClassOf ? f)]</pre>
典型案例库 获得分级结果	<pre>[rule2:(? a rdfs:subClassOf ? b)(? b owl:onProperty case:属于平台)(? b owl:someValuesFrom ? c) (? c rdfs:subClassOf ? d)(? d rdfs:subClassOf ? e)(? e owl:onProperty case:平台内容为) (? e owl:someValuesFrom ? f)(? g rdf:type owl:Restriction)(? g owl:onProperty case:应) (? g owl:allValuesFrom ? f)-&gt;(? a rdfs:subClassOf ? g)]</pre>



增加,存在条款冲突的情况,给出了冲突检测的推理规则,若两条条款反映的是同一关键词,但两条条款涉及的分级结果不一致,则两条条款冲突,此时效力较低的政策法规服从效力较高的政策法规,分级以效力较高的政策法规作为分级依据。表中还给出了获得分级结果的推理规则,若某条款反映某一关键词,条款涉及某个分级内容(这里以无条件开放为例),则涉及这一关键词的领域数据应当无条件开放;若某案例属于某一平台,此平台涉及某个分级内容(这里以无条件开放为例),则此案例应当无条件开放。

### 2.4 查询实现

文中基于自定义规则,采用 SPARQL 查询语句实现推理查询功能<sup>[14,28]</sup>。对查询推理的每个输入词进行分级结果查询,输出与输入词相关的政策法规条例,并检测是否有与条例相冲突的其他条例;同时根据政策法规条例所属类别,按其效力进行从高到低的排序,

并选取效力最高的结果作为每个输入词对应的中间结果;若在政策法规库中查找不到结果,则去典型案例库中查找,将输入词与案例所属平台、案例名称作为参考依据给出。

根据上述中间结果,开放以不与开放、依申请开放、无条件开放的从高到低的级别等级,共享以不予共享、有条件共享、无条件共享的级别等级,输出开放和共享最高等级的结果,并输出所有对应的法律法规条例作为参考依据。

### 3 系统实现与结果评估

图 6 为所创建的 GSGD 系统,输入拟分级数据后,上述模块会对数据进行计算、推理、分析,最终系统会自动给出分级结果及其依据,点击依据条例,系统会显示详细的条例信息。



图 6 系统测试示例

为验证所实现系统的效果,文中采用欧氏距离(Euclidean distance)作为相似度计算对比方法进行实验。实验数据为 xx 市若干委办局“行政区 委办局名称 系统名称 表名称 字段名称”目录,共 500 条,涉及卫计委、国税局、城管局、公安局等委办局数据目录;由

于数据是无标签的,因此对数据分别从开放与共享两个方向进行人工标注,以方便对比实验结果。实验结果也分别从开放与共享两个方向进行对比,由表 3 可看出,不论是开放还是共享方向,文中方法相比于对比方法在准确率、F<sub>1</sub>值上更高,验证了该方法的有效性。

表 3 两种方法对比结果(对共享、开放方向进行分级)

方法	共享方向分级结果			开放方向分级结果		
	P / %	R / %	F <sub>1</sub> / %	P / %	R / %	F <sub>1</sub> / %
文中方法-余弦相似度	71.2	71	61	77.2	77	74
对比方法-欧氏距离	70.6	71	59	76.6	77	68

### 4 结束语

针对政府数据分级工作数据资源规模大,支撑依据不足、主观性强、精确性差等问题,提出了采用政策法规库以及典型案例库对数据进行自动化分级,设计

并实现了基于 BERT 以及本体构建推理的政府数据分级系统—GSGD。通过 BERT 以及相似度计算获取本体推理查询的输入关键词,再通过 Jena 推理机进行推理查询,实现对政策法规冲突检测、效力级别分析等功能,最终获得分级结果以及依据;最后通过对比实验分

析,验证了该方法的有效性。未来在以下几个方向有待探索:一、采用人工构建本体,但随着政策法规/案例的增加,应尝试采用自动化方法构建政策法规库以及案例库;二、调整相似度计算方法,将多种相似度计算方法融合以得到更精确的结果。

#### 参考文献:

- [1] 焦海洋. 中国政府数据开放共享的正当性辨析[J]. 电子政务, 2017(5): 19-27.
- [2] 朱 琪. 详解十三五: 推进数据资源分类、分级管理[EB/OL]. [2019-09-24]. [http://china.cnr.cn/ygxw/20160621/t20160621\\_522454403.shtml](http://china.cnr.cn/ygxw/20160621/t20160621_522454403.shtml).
- [3] VALENTE A, BREUKER J. A functional ontology of law [R]. Padua, Italy: Cdam Publishers, 1994.
- [4] BREUKER J, VALENTE A, WINKELS R. Legal ontologies in knowledge engineering and information management[J]. Artificial Intelligence and Law, 2004, 12(4): 241-277.
- [5] 何 庆, 汤 庸, 黄永钊. 基于本体的法律知识库的研究与实现[J]. 计算机科学, 2007, 34(2): 175-177.
- [6] 贾君枝, 郭丹丹. 法律框架网络知识本体构建与实现[J]. 情报学报, 2007, 26(5): 733-740.
- [7] 卢明纯. 基于 OWL 本体的法律知识库原型系统的设计和实现[J]. 现代情报, 2010, 30(7): 34-38.
- [8] 余贵清, 张永安. 基于本体的刑事审判案例推理方案研究[J]. 图书情报工作, 2014, 58(13): 118-124.
- [9] 姜 赢, 张 婧, 朱玲萱. 基于本体的医疗卫生政策法律知识管理系统[J]. 中华医学图书情报杂志, 2016, 25(12): 11-17.
- [10] THAMMABOOSADEE S, KIATTISIN S, DARAKORN S, et al. Sentence identification system based on criminal law ontology[J]. International Review of Law, Computers & Technology, 2017, 31(3): 308-322.
- [11] 贾君枝, 毛海飞. 基于法律框架网络本体的语义匹配技术研究[J]. 情报理论与实践, 2008, 31(1): 124-128.
- [12] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics: human language technologies, NAACL-HLT. Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.
- [13] 梁 晔, 刘宏哲. 运用 Jena 对本体模型进行推理及其应用[J]. 北京联合大学学报: 自然科学版, 2009, 23(3): 23-27.
- [14] 希茨利尔. 语义 Web 技术基础[M]. 俞 勇, 译. 北京: 清华大学出版社, 2012: 184-211.
- [15] 杜文华, 董 慧. 本体建设工具比较研究[J]. 情报杂志, 2005, 24(2): 5-7.
- [16] 孙 瑾. 本体编辑工具的分析与研究——Protégé2000 对中文本体编辑的适用性探析[J]. 图书情报工作, 2006, 50(12): 26-29.
- [17] 赵忠君. 土地法律本体构建及其推理机制研究[D]. 武汉: 武汉大学, 2011.
- [18] 蔡鑫怡, 姜威宇, 韩浪焜, 等. Bert 在中文阅读理解问答中的应用方法[J]. 信息与电脑, 2019(8): 39-40.
- [19] 俞敬松, 魏 一, 张永伟. 基于 BERT 的古文断句研究与应用[J]. 中文信息学报, 2019, 33(11): 57-63.
- [20] YAO L, JIN Z, MAO C, et al. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora[J]. Journal of the American Medical Informatics Association, 2019, 26(12): 1632-1636.
- [21] 吴 旭, 郭芳毓, 顾夏青, 等. 面向机构知识库结构化数据的文本相似度评价算法[J]. 信息网络安全, 2015(5): 16-20.
- [22] GUNAWAN D, SEMBIRING C A, BUDIMAN M A. The implementation of cosine similarity to calculate text relevance between two documents[J]. Journal of Physics: Conference Series, 2018, 978(1): 012120.
- [23] 宗成庆. 统计自然语言处理[M]. 第 2 版. 北京: 清华大学出版社, 2013: 418-437.
- [24] 潘 超, 古 辉. 本体推理机及应用[J]. 计算机系统应用, 2010, 19(9): 163-167.
- [25] 纪兆辉. 本体的推理研究[J]. 南京师范大学学报: 工程技术版, 2012, 12(3): 54-59.
- [26] 韩 昊, 李禹生. Jena 智能推理查询中的自定义规则构造方法研究与应用[J]. 软件导刊, 2014, 13(7): 13-15.
- [27] 陈和平, 郭晶晶, 吴怀宇, 等. 基于 Ontology 和 Jena 的个性化 E-Learning 系统研究[J]. 武汉理工大学学报: 交通科学与工程版, 2007, 31(6): 1049-1052.
- [28] ALLEMANG D, HENDLER J. 语义万维网: 工程实践指南[M]. 张自力, 李 莉, 译. 第 2 版. 北京: 高等教育出版社, 2015: 54-110.