基于知识图谱的政府结构化库表数据自动分类算法研究

王胜漪[1,2] 刘汪洋[1,2] 邹佳[1,2] 蔡惠民[1,2]

1中电科大数据研究院有限公司，贵州 贵阳 550081
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 550081
wangshengyi@cetcbigdata.com

提升政府治理能力大数据应用技术国家工程实验室
中电科大数据研究院有限公司

CCKS 2020

摘要

政府数据资源共享开放是大数据产业发展的重要途径，政府数据分类有利于实现数据高效管理，是实施各类政府数据共享开放策略的前提。本文针对政府结构化库表数据分类任务中人工标注成本高、自动分类准确率低等问题，提出了基于知识图谱的政府数据自动分类算法。以政务领域知识图谱为核心，结合BERT扩展数据训练模型、基于知识图谱单主题分类、基于TF-IDF多主题权重分类方法共同实现零标注政府数据分类。经实验表明，结合知识图谱的自动分类方法表现出良好的分类效果，为知识图谱在政府数据的应用开拓了新领域，促进政府数据资源的共享开放。

研究背景

在大数据发展的今天，政府数据资源的开放共享是大数据产业蓬勃发展的关键，然而，我国分类管理体系的不健全导致政府数据开放共享进程受到严重阻碍。数据的分类管理有助于理清数据管理和共享开放的义务及权利，帮助政府加快推动政务信息系统互联及数据共享，增强政府公信力，为大数据产业发展提供安全支撑。

算法介绍

整个分类算法可划分为四个模块，分词模块采用jieba分词，同时添加政府领域核心词汇库，将输入数据划分为单个词组；KG-BERT模块是结合政务知识图谱和BERT模型共同构建的政府数据分类模型，主要针对政务知识图谱中查询无结果的库表数据进行分类；KG-STopic模块是将政务知识图谱筛选出的单主题查询输出直接作为最终分类结果；KG-MTopic模块是结合源词扩展和TF-IDF实现数据的主题权重判定，通过该模块实现多主题数据的类别划分。

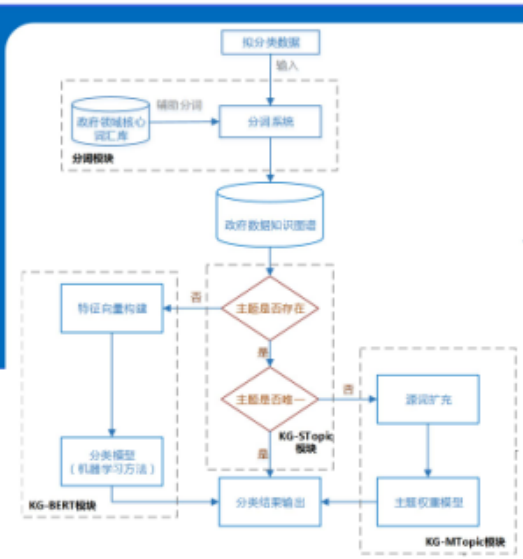

图2 KG-BERT算法流程图

1. KG-BERT

结合政府领域核心词库完成jieba分词，划分核心词汇；通过知识图谱，查询核心词的唯一主题词，得到ST数据集；通过匹配规则修正数据，得到PM数据集；通过结合政务知识图谱和LDA模型实现数据扩展，得到KE数据集；将该数据集用于训练BERT模型，完成分类任务；

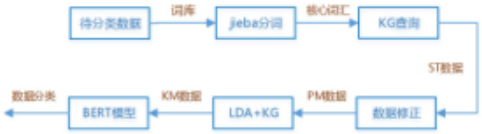

图1 本文算法流程图

2. KG-Mtopic

对主题词对应的关键词进行扩展；将分词后的结果再一次输入政务知识图谱进行主题查询；由原始主题和拓展后的主题计算出每个原始主题的TF-IDF权重；对各原始主题的TF-IDF权重进行排序，选取权重最大主题作为原数据类别；

$$TF = \frac{\text{扩展后的主题与某原始主题相同的个数}}{\text{扩展后的主题数}}$$

$$IDF = \log \frac{100}{\text{某原始主题下经扩展得到的主题数}}$$

图3 KG-Mtopic流程图

实验结果

实验数据均来自于某地级市政府“各委办局-系统-表-字段”全量目录，共计106698条无标签数据。为验证算法准确率，本文抽取十个类别中1500条库表数据作为测试数据。

表1 某地级市政府数据样例表

序号	测试数据	分类标签
1	××市 工信委 一企一策 企业基本信息-激活人手机号	企业
2	××市 司法局 律师管理信息系统 案件备案 案由	法院
3	××市 卫计委 ××市全员人口服务管理信息系统 现有夫妇情况查询-选法生育夫妇名册 乡	计划生育
4	××市 交管局 ××省道路交通安全综合监管云平台 驾驶人户籍化管理-职业信息 员工编号	交通管理
5	××市 卫计委 结核病管理信息系统 快速诊断附药结核患者实验室检查情况 快速诊断结果人数	医药卫生管理

本文方法和无扩展结合算法同KG-BERT扩展分类算法结果相比，其精确率、召回率和F1值均有一定提升并且，前两种算法在十个类别数据中有相似的高精度曲线，这说明多模块的混合策略更适用于政府结构化库表数据的分类。从实验结果可看出，以精确率、召回率和F1值作为对比指标，本文提出的算法在政府结构化库表数据分类任务中表现最佳，证明本文KG-BERT模块在扩展数据下表现出更好的分类结果，在无标注数据训练模型的前提下实现了良好的政府数据分类。

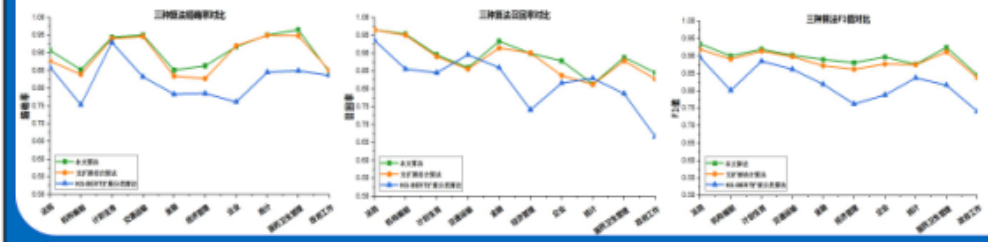

图4 分类平台演示图

结论

本文采用多模块结合的方法实现政府结构化库表数据分类。利用KG-STopic模块结果作为BERT模型训练集，为了防止模型过拟合，提出KG-BERT方法，通过政务知识图谱实现数据扩展。然后，针对多主题数据提出基于TF-IDF的主题权重分类方法。最后，结合KG-BERT、KG-STopic、KG-MTopic模型共同实现政府结构化库表数据自动分类。经实验表明，本文提出的方法在政府数据分类任务中表现较好，可为政府数据资源开放共享提供重要技术支撑。