

构建面向实际应用的科技文献自动分类引擎^{*}

张智雄 赵 旻 刘 欢

摘 要 文献分类是图书馆情报学领域的一个传统研究问题。实用化的中图法自动分类系统最重要的一个要求就是能够将文献精确地自动分类到三级或四级类目之下,这意味着需要将特定文献较为精确地自动分类到上千个类目之下。为了构建面向实际应用的科技文献中图法自动分类引擎,本文基于层次分类思想,设计和实现了一个基于多层分类器集群的科技文献自动分类引擎系统,并重点解决了科技文献自动分类引擎建设中的四个关键问题:①如何获取并构建大规模高质量分类训练数据以提升自动分类效果;②如何设计和实现多层分类器集群以有效解决上千个类目自动分类的准确性;③如何面向现实要求来优化处理流程以提升分类速度;④如何设计和开放接口以支撑引擎的开放调用。最终构建了科技文献自动分类引擎,各项指标达到了实用化要求,初步实现了基于中图法的自动分类系统的实际应用。图4。表7。参考文献16。

关键词 文献分类 科技文献 自动分类 分类引擎 层次分类法 分类器集群

分类号 G254.29 TP391

Construction of a Practical Application-oriented Automatic Classification Engine for Scientific Literature

ZHANG Zhixiong, ZHAO Yang & LIU Huan

ABSTRACT

Literature classification is a traditional research problem in library and information science. One of the most important requirements for a practical automatic classification system based on Chinese Library Classification is to automatically classify literature into its third or fourth level. This means that it is necessary to be able to accurately and automatically classify specific literature into thousands of categories of Chinese Library Classification. In order to build a practical application-oriented automatic classification engine of scientific literature, we use the idea of hierarchical classification to design the engine's logical and technical framework, and realize an automatic classification engine system for scientific documents based on a multi-layer classifier cluster.

This paper focuses on four key issues in the construction of automatic classification engine for scientific literature: 1) How to obtain and construct large-scale high-quality classification training data to improve the effect of automatic classification; 2) How to design and implement a multi-layer

^{*} 本文系中国科学院文献情报能力建设专项课题“基于科技文献知识的人工智能(AI)引擎建设”(编号:E0290906)的研究成果之一。(The article is an outcome of the special sub-project “Artificial Intelligence (AI) Engine Construction Based on Scientific Literature Knowledge” (No. E0290906) supported by the Chinese Academy of Sciences' Literature and Information Capacity Building.)

通信作者:张智雄,Email:zhangzhx@mail.las.ac.cn, OCRID:0000-0003-1596-7487 (Correspondence should be addressed to ZHANG Zhixiong, Email:zhangzhx@mail.las.ac.cn, OCRID:0000-0003-1596-7487)

classifier cluster to effectively solve the accuracy problem of thousands of categories' automatic classification; 3) How to face practical requirements and optimize the processing flow to improve the classification speed; 4) How to design and open the interface to support the open call of the engine. In view of the above four key issues, we propose corresponding solutions: 1) Building large-scale high-quality classification training data based on the Chinese Science Citation Database (CSCD) to improve the classification effect. In addition, we design category screening rules to balance the unevenness of categories among all levels; 2) We construct 26 automatic classifiers based on the BERT pre-training language model, and assemble them into a two-layer automatic classification cluster framework to achieve accurate automatic classification of scientific literature; 3) Compressing the classification model based on the microservice, and optimizing the batch literature classification processing flow to improve classification speed; 4) Developing an HTTP-based Application Programming Interface (API) so that the classification engine can be openly called.

The study finally built an automatic classification engine for scientific literature that includes 2118 categories and a total of about 1.8 million high-quality scientific literature corpus, and it can automatically classify literature into the third or fourth level of the Chinese Library Classification. Moreover, we select journal papers and dissertations to evaluate the classification effect, and the complete discrepancy rate between the classification results and the original classification number is only about 1%. In addition to the classification effect, the classification speed and other indicators of the classification engine have all reached the practical requirements, which proves that the practical-level automatic classification system based on the Chinese Library Classification method has been initially realized. 4 figs. 7 tabs. 16 refs.

KEY WORDS

Literature classification. Scientific literature. Automatic classification. Classification engine.
Hierarchical classification. Classifier cluster.

0 引言

文献分类是图书馆学情报学领域的一个传统研究问题。目前将文献分到几个或几十个类下的自动分类系统已较为常见,但这并不能满足实际应用的需求。《中国图书馆分类法》(以下简称“中图法”)是我国普遍使用的综合性分类法。为了实现中图法自动分类系统的实用化,需要将文献精确地分类到中图法的三、四级目录下。经统计,中图法的三级类目共计约1700个,这就意味着自动分类系统需要有效地将特定文献较为精确地自动分类到上千个目录下,只有如此才能够满足现实的需要。

关于成百上千个类目的自动分类问题,已

经有学者采用层次分类法展开了研究^[1]。层次分类法是根据类目之间的层次关系,将一个复杂的分类任务分解成为若干个规模较小的局部分类任务的一种方法^[2]。它首先对输入的文献进行高层自动分类以获得高层分类号,之后再将这一文献在高层分类号下继续进行下层类目的自动分类。这种方法能将复杂问题简单化,不仅极大降低了计算复杂度,而且适用于规模较大、具有层级结构分类体系的自动分类任务。

近年来,一些学者尝试利用层次分类法对中图法的层级分类体系进行自动分类研究。王昊等^[3]利用层次分类理念将中图法转化为三层分类体系,共选取17个类目对期刊论文进行自动分类研究。刘高军等^[4]基于极

限学习机(ELM)和混合特征构建了面向中图法的层次分类器,共选取了13个类目进行试验性研究。冉亚鑫等^[1]提出了基于Stacking算法的层次文本分类方法,选取了中文期刊数据库约73万篇文献、773个三级类目进行层次分类研究。

上述研究成果表明,基于层次分类的中图法自动分类是可行的,这些探索性研究对本文具有启示作用。同时也应认识到,上述研究均只构建了实验模型,未形成可供开放测试的系统,离实际应用还有很大的距离。因此,如何将文献快速、较为准确地分类到上千个类目下,这一问题仍然没有得到很好的解决。

为有效解决这一问题,本文构建了一个精准、快速、可支持开放调用的科技文献中图法自动分类引擎。这一分类引擎基于层次分类思想,以多层分类器集群为主体框架,重点解决了四个关键技术问题:①构建大规模高质量分类训练数据,以提升自动分类效果;②基于层次分类思想构建多层分类器集群,以实现上千个类目的准确分类;③优化改进分类流程,提升分类速度,以满足现实需要;④制定API接口规范,以支撑引擎的开放调用。最终构建了一个能够将科技文献自动分类到中图法三、四级类目下的科技文献自动分类引擎,实际测试效果显示,各项指标基本达到了实用化要求,初步实现了基于中图法的自动分类系统的实际应用。

本文重点论述面向实用化的科技文献分类引擎的整体架构,并对引擎建设中的四个关键问题及解决方案、系统实际实现及效果测评进行阐述。

1 科技文献自动分类引擎的整体架构

如前所述,笔者基于层次分类思想来实现科技文献中图法自动分类引擎的构建。在开展大规模分类引擎构建之前,已经通过实验^[5]证实了这一方法的有效性和可行性。

具体而言,笔者构造了两层分类器,将医学文献自动分类到中图法医学的三级类目下。第一层分类器将医学文献自动分类到中图法医学的16个二级类目下,第二层分类器由16个分类器组成,根据第一层分类的结果进一步将文献自动分类到各自的三级类目下,总计实现了112个三级类目的分类。实验结果表明,两层分类的分类结果比直接利用112个类目进行单层分类的效果有了显著提升,F1值比单层分类提高了4.39%。

面向实际应用的科技文献中图法自动分类引擎的建设,需要在此实验的基础之上,从医学领域扩展到中图法的全部领域,引擎的类目数也从数百个扩展到数千个。科技文献自动分类引擎的整体架构设计如下。

1.1 科技文献自动分类引擎的逻辑框架

在层次分类引擎的逻辑框架构建中,首先需要解决的是分类引擎的层数问题。理想情况下,可以按中图法的类目层数来决定分类引擎的层数。但分类引擎层数的增加意味着需要更多的分类器,这将会使整个分类引擎的计算复杂度增加,资源消耗巨大,同时整体的分类速率也将受到影响。

基于上述原因,本文没有采用按中图法的类目层数来决定分类引擎层数的方案,而是采用了两层的层次分类体系方案,即基于两层分类器集群构建支持科技文献自动分类引擎的逻辑框架。

具体而言,第一层分类器(根分类器)主要解决中图法中一级类目的分类;第二层分类器为多个领域分类器集群的组合,主要针对各个大类下具体的细粒度类目进行分类。

当需要分类一篇文献时,首先经过第一层的根分类器将其分类到中图法的一级类目之下,随后进入专门针对该一级类目构建的第二层领域分类器对文献进行更加细粒度的分类,最终输出该文献的具体分类号。分类器集群设计的主要思路如图1所示。

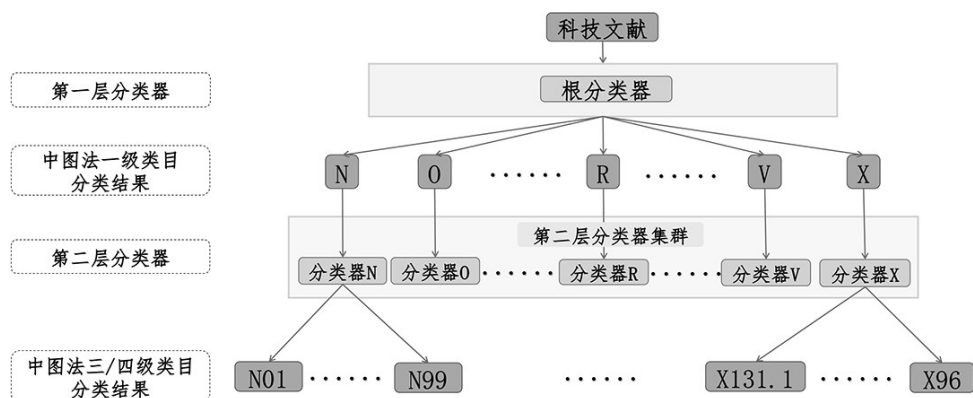


图1 支撑上千个类目的中图法两层分类器集群设计思路

1.2 科技文献自动分类引擎的技术框架

科技文献自动分类引擎需要有效组合计算能力、训练数据、分类模型、调用接口等多项功能。基于多层分类器集群的科技文献自动分类引擎的技术架构如图2所示。

其中,高性能计算层通过包括图形处理器(Graphics Processing Unit, GPU)在内的计算资源为引擎提供强大的计算能力。多层分类器集群中包括了多个自动分类器,而每个自动分类器都以高质量的大规模文献分类数据为基础,

利用深度学习方法,实现了自动分类模型的有效训练。系统通过API接口来实现引擎的开放调用。

用户可以利用客户端软件、浏览器或编写的程序代码等,来调用引擎的API接口。客户端将需要分类的文献提交给自动分类引擎,自动分类引擎在接收到输入的文献之后,通过多层分类器集群实现文献的自动分类,并将文献的自动分类结果返回给客户端。

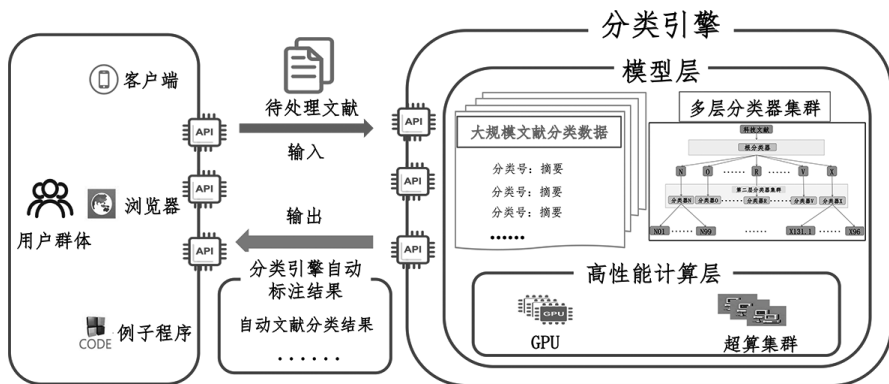


图2 基于多层分类器集群的科技文献自动分类引擎技术架构

2 科技文献自动分类引擎建设的关键技术问题及解决方案

基于上述框架构建的科技文献自动分类引

擎,尚有四个关键技术问题需要解决:①如何获得大规模高质量分类训练数据以提升分类效果;②如何构建多层分类器集群以实现精准分类;③如何优化分类器集群的处理流程以提升分类速度;④如何制定接口规范以支撑引擎的

开放调用。

针对上述四个关键问题,本文提出了相应的解决方案:①基于中国科学引文数据库(CSCD)构建大规模的高质量分类训练数据以提升分类效果;②基于BERT预训练语言模型来构建多层分类器集群以实现精准分类;③基于微服务和处理流程优化以提升分类速度;④制定基于HTTP的应用程序编程接口(API),以使分类引擎能够被开放调用。

2.1 基于CSCD构建大规模高质量分类训练数据以提升分类效果

训练数据的规模和质量对深度学习模型的效果有重要影响。构建大规模高质量的自动分类训练数据能够有效提升科技文献自动分类引擎的分类效果。因此,以中国科学院文献情报中心研发的中国科学引文数据库(CSCD)为基础,通过筛选得到高质量分类语料并确定分类类目,解决了不同类目下文献数据分布不均问题,最终构建了包括2118个类目、总计约180万篇高质量科技文献的自动分类训练数据集。

2.1.1 筛选高质量分类训练数据

为筛选高质量科技文献分类训练数据,需要解决的首要问题是如何选择与科技文献相对

应的中图法分类号。首先选择中图法中与自然科学相关的10个一级类目(N—X)作为第一层分类器的类目。其中,由于工业技术(T)类目涵盖的学科领域宽泛,其下的二级类目代表了不同领域的工业技术分类,区分性较强,且各个二级类目下的文献数据量较大,故将T类目下的16个二级类目(TB—TV)调整至第一层分类器的类目中。第一层分类器的类目数量最终设定为25个。

中国科学引文数据库是一个高质量的文献数据库,其收录的文献都有相应的中图法分类号。这些文献的分类号有的由作者提供,有的由数据库建设者标引。文献数据的基本结构如图3所示。

以CSCD高质量文献数据为基础,对第一层分类器25个类目的分类训练数据进行进一步筛选,确保每篇文献与其对应类目的唯一性。根据文本分类任务中“分类号—科技文献摘要”的训练数据格式,抽取CSCD中文科技期刊论文的分类号和中文摘要两个字段,将包含两个及以上分类号的文献数据删除,保留仅包含一个分类号的文献数据。经筛选后共计得到2236826条分类训练数据。

classification	title	abstract
TU533	石墨铜珠复相智能环氧树脂混凝土的拉...	土木工程中构件受拉区的变形和裂缝影响其承载力和正常使用,因此研究受拉构件的传感材料有重...
TU528	高温环境下硫磺砂浆-钙矾石固溶体...	由硫磺砂浆引起的硫酸盐侵蚀破坏在许多国家都有相关工程案例。铝相以加速硫磺砂浆的生...
TU528.45	灰粉比对苯丙乳液基水泥复合填缝料研...	制备了一种适用于机场水泥混凝土路面接缝的苯丙乳液基水泥复合填缝料(SCS),采用HS-3001B型研...
TU528	模拟酸雨环境下废瓷砖再生混凝土中性...	将废弃瓷砖破碎为再生粗、细骨料及粉料,全组分制备再生混凝土。采用室内模拟酸雨侵蚀试验,研...
X703	磁性羟基磷灰石的制备及其除氟性能研究	氟离子在饮用水中浓度超过1.0mg/L,将对人体健康造成极大危害。通过原位共沉淀法将具有磁性...
P223	应用Slepian局部谱方法估算中国大陆...	根据经典的球谐函数方法,为满足正交化要求,观测数据需要覆盖整个球面,而对于地表局部测量数...
P237	高分辨率遥感影像的深度学习变化检测...	为提升高分辨率遥感影像的变化检测精度,提出一种利用深度学习的变化检测方法。在预处理的基...
P227	逐星25号无场化相对辐射定标	逐星25号是中国首颗高轨亚米级高分辨率光学逐星卫星。常规的偏航相对辐射定标利用卫星或相...
P237	多源遥感影像湿地检测概率潜在语义分析	提出了一种基于概率潜在语义分析的多源遥感影像湿地检测方法。首先提取高分辨率影像的光谱、...
TQ174	Al ₂ O ₃ 溶化法制备微细Al ₂ O ₃ -3 /Al复...	真空条件下,以Al ₂ O ₃ -3和Al为原料,通过Al ₂ O ₃ 溶化法制备微细Al ₂ O ₃ -3/Al复合粉体。XRD和SEM分析...
TQ172	硫石膏-水泥-矿粉复合材料的性能研究	硫石膏-水泥-矿粉复合材料的水泥和矿粉,组成硫石膏-水泥-矿粉复合材料,主要研究了其耐...
TU528	显微硬度分析在再生混凝土多界面研究...	与普通混凝土相比,再生混凝土存在多薄弱界面结构,该薄弱结构是再生混凝土性能低下的主要原...
U416.216	基于CPM模型的船用混凝土组成设计方法	针对基于三重球模型的混凝土组成设计方法存在参数计算复杂,在实际工程中使用不便,对其抗压...
TU308.1	混合碳纤维活性粉末混凝土力学性能研究	采用两种不同尺寸的碳纤维混合掺入活性粉末混凝土中,通过抗压、劈裂和四点弯曲的力学性能试...
TU528	活性粉末混凝土抗压强度试验研究	通过立方体抗压强度试验研究了活性粉末混凝土的受压破坏过程,并通过单因素对比试验研究了水...
TQ126.3	钾长石作为助熔剂对磷石膏还原反应的可行...	研究了钾长石对磷石膏还原反应的工艺及动力学,考察了反应温度、反应时间、酸度值及焦炭过...
U414.75	沥青路面用复合变形相变材料的耐用性...	为研究复合变形相变材料对沥青混合料耐用性能的影响,制备沥青路面用复合变形相变材料,测定相...
U416.217	偶联剂改性沥青与碎石集料界面性能...	为了改善酸性碎石集料与沥青粘附性,利用硅烷偶联剂KH550对沥青进行改性。采用针入度、延度...
TU528	不同温度下混凝土补强在不同温度下的力学性能	本研究主要探讨混凝土补强在不同温度下的力学性能,以普通混凝土添加不同碳纤维含量的活性粉...
TB321	高温发泡法制备粉煤灰隔热耐火砖	以高铝粉煤灰为主要原料,高铝铝铁合金渣为辅助原料,采用高温发泡法制备一种新型粉煤灰隔热...
TU411.8	硅烷改性聚脲酸酯-水-Na ₂ O ₂ -水...	采用γ-甲基丙稀氧基丙基三甲氧基硅烷(KH570),丙稀酸(AA)以及甲基丙稀氧基乙醚(MPEO)...
TU411.8	锌-乳抗面白斑黑粉体的改性及抗菌性...	通过溶胶-凝胶法制备了锌-乳抗面白斑黑粉体。通过添加表面活性剂对材料进行改性,并探究了材料改...
TQ172.1	快凝快硬高贝利特硫磺砂浆水泥熟料水...	研究了一种新型快凝快硬高贝利特硫磺砂浆水泥的水化性能,并利用量热仪、XRD、TGA、SEM等方...
TU528	碱矿渣泡沫混凝土的吸声性能试验研究	为了研究开发一种新型的绿色环保吸声材料,以化学发泡法制备碱矿渣泡沫混凝土试样,通过实验分...
TQ172	硅质土的水灰比活性试验研究	采用对比强度法与结合水率法,分别研究了机械活化、有氧与无氧活化对硅质土火山灰活性的影...
TU528	硅粉抑制碱集料反应的试验研究	通过对青海地区的集料调研分析,发现青海大部分地区的集料存在高的碱活性。采用西宁城北地区...
TQ646	石煤提钼尾矿制备硫酸盐水泥熟料的研究	以石煤提钼尾矿、硫酸钙、粘土和铁粉为原料,在不同煅烧条件下制备了不同配比的产物,并研...

图3 中国科学引文数据库(CSCD)文献数据基本结构

2.1.2 解决不同类目下文献数据分布不均问题

训练数据的均衡情况对于分类效果具有重要影响。经统计,上述约 224 万条文献分类数据共包含 10 148 个类目,这些文献在第一层分类器 25 个类目下的数量分布并不均匀。为此,笔者制定了分类训练数据的类目筛选规则,以解决不同类目下文献数据分布不均问题。具体规则如下。

(1) 以设定的 25 个第一层分类器的类目为起点,对各个类目按照中图法的层次由上至下逐层进行下级类目的文献数据量筛选统计工作。

(2) 若某个类目的名称不是很明确(如: R723.19 其他),则将该类目下的文献数据合并

到其上级类目中。

(3) 若某个类目下的文献数据量小于 100 条,则将该类目下的文献数据合并到其上级类目中。

(4) 若某个类目下的文献数据量在 100—2 000 条之间,则将该类目列入最终的类目列表中,不再继续统计其下级类目的文献数据量。

(5) 若某个类目下的文献数据量大于 2 000 条,则继续统计其下级类目的文献数据量,一直统计至各个大类的四级类目或五级类目(视各类目下的数据量情况而定),将该四级类目或五级类目列入最终的类目列表中。

最终共计筛选得到 2 118 个类目,整个类目筛选流程如图 4 所示。

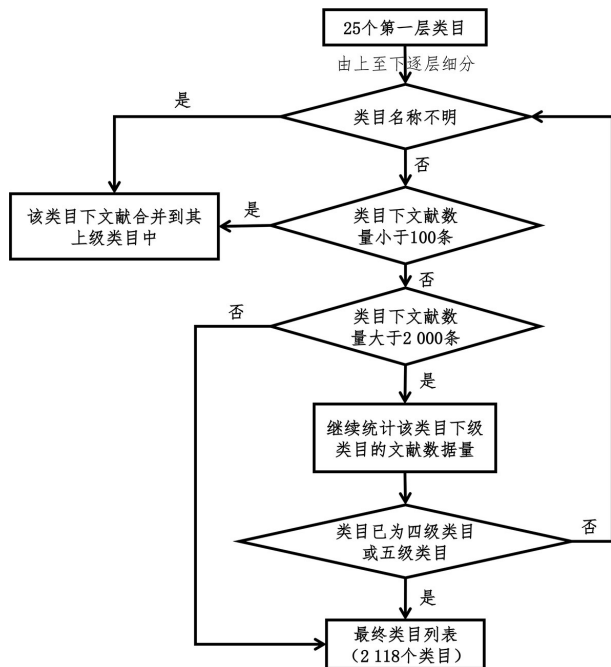


图 4 类目筛选流程

以“P58 岩石学”的下级类目筛选统计情况为例说明上述规则。在 P58 下级类目的筛选统计之前,需先经过 P—P5—P58 的类目筛选统计过程。P58 类目下的文献数量为 2 712 条,因其文献数据量超过 2 000 条,所以继续统计其所有下级类目的文献数量。P58 的下级类目共有 8 个,其名称及对

应类目下的文献数量如表 1 所示。根据类目筛选规则,P581 和 P588 类目下的文献数量因超过了 100 条予以保留,故在 P58 中相应地去掉 P581 和 P588,再将其他文献数量不超过 100 条的下级类目合并到其上级类目 P58 中,共计获得 280 条摘要数据,记作 P58 * (P581,P588)。

表1 “P58 岩石学”的下级类目及相应文献数量

类 目	文献数量
P581 岩石成因	241
P583 岩石产状、结构和构造	36
P584 岩石物理与岩石化学	99
P585 岩石鉴定、分析	42
P586 岩相学	17
P587 岩石分布、区域岩石志	14
P588 岩石分类	2 191
P589 实验岩石学、工艺岩石学	33
其他不属于上述类目的不规范类目	39

由于 P588 类目下的文献数量超过了 2 000 条,所以继续统计其所有下级类目的文献数量。P588 下共有三个下级类目,这三个类目下的文献数量均超过 100 条,可以全部保留,同时这三个类目下的文献数量均不超过 2 000 条,故无需再向下细分,最后删除上级类目 P588,经筛选后的 P58 下级类目和文献数量如表 2 所示。其他类目的下级类目筛选统计工作亦是如此。

表2 筛选后的“P58 岩石学”下级类目及相应文献数量

类 目	文献数量
P58 * (P581, P588) 岩石学	280
P581 岩石成因	241
P588.1 岩浆岩(火成岩)	1 186
P588.2 沉积岩	661
P588.3 变质岩	259

本文以设定的 25 个类目为基准,利用上述规则进行各类目下级类目的筛选统计,最终构建起了包括 2 118 个类目、总计 1 794 428 篇高质量科技文献的自动分类训练数据集。具体分布情况如表 3 所示。

表3 训练数据的类目和文献总量统计

类 目	类目总量	文献总量
N 自然科学总论	13	3 286
O 数理科学和化学	108	109 920
P 天文学、地球科学	147	104 002
Q 生物科学	116	67 109
R 医药、卫生	482	448 534
S 农业科学	276	176 659
TB 一般工业技术	28	32 017
TD 矿业工程	21	10 956
TE 石油、天然气工业	42	32 241
TF 冶金工业	12	8 257
TG 金属学与金属工艺	94	62 891
TH 机械、仪表工业	46	32 970
TJ 武器工业	15	6 834
TK 能源与动力工程	22	25 136
TL 原子能技术	18	9 744
TM 电工技术	69	59 720
TN 无线电电子学、电信技术	101	78 253
TP 自动化技术、计算机技术	40	181 974
TQ 化学工业	88	57 747
TS 轻工业、手工业	47	48 523
TU 建筑科学	95	61 865
TV 水利工程	32	15 885
U 交通运输	79	41 632
V 航空、航天	49	27 652
X 环境科学、安全科学	78	90 621
总计	2 118	1 794 428

2.2 基于 BERT 模型构建多层分类器集群以实现精准分类

基于自动分类引擎的逻辑框架,笔者利用 BERT 模型构建了 26 个自动分类器,并将其组装入一个两层的自动分类集群框架之中,以实现科技文献的自动分类。

BERT 语言模型采用双向编码,结合不同语境动态获取词语在上下文中的不同语义特征,其特征获取效果比普通模型要强^[6]。很多学者通过微调 BERT 来丰富特征表示^[7-11]。对于科

技文献摘要来说,BERT 模型可以很好地学习整个摘要文本内容的语义特征。通过在构建好的大规模高质量训练数据上进行微调,能够取得比传统机器学习方法更好的分类效果。

通过微调 BERT 模型对第一层分类器及第二层的各个领域分类器进行训练。对于第一层分类器的训练,类别数对应设定为 25 个,训练样本的数量为 1 794 428。根据第一层分类器的类目,采用同样的方法训练各个类目下对应的第二层领域分类器,第二层的分类器集群共包括 25 个领域分类器,各个领域分类器的类目数量及训练文献数量如表 3 所示。

将两层共计 26 个分类器根据第一层分类器的 25 个类目相映射即构成了支撑上千个类目的两层分类器集群。具体而言,在对科技文献进行自动分类时,科技文献分类引擎首先调用第一层分类器对输入的文献摘要进行分类,得到第一层 25 个类目的分类结果。引擎根据预测结果的分值和预设的阈值进行判断,取得分最高的 1—2 个类目。根据第一层分类器的分类结果,引擎调用与该结果相对应的第二层领域分类器,对输入的文献进行再次分类。同样根据预测结果的分值和预设的阈值进行判断,取得分最高的 1—2 个类目作为最终的分类结果。根据模型预测得分,每篇文献最多可得到 4 个分类号。

2.3 基于微服务的处理流程优化以提升分类速度

科技文献自动分类引擎在面向实际应用时,需要能够实时响应分类请求。经实测,在单个 Nvidia GTX 1080Ti GPU 的硬件环境中加载 BERT-Base 预训练模型的平均时间在 30 秒以上,利用两层分类器集群进行分类时需分两次

调用分类器,由此利用 Google 提供的基于 TensorFlow 框架的原始代码进行单篇文档的分类时间将达 60 秒以上。这样的速度远远无法满足对大规模科技文献数据进行自动分类的需求。

针对此种情况,采用基于微服务的处理流程优化的思路方法,有效地提升了分类速度。具体方法如下。

(1) 压缩模型以减少载入时间。利用 FREEZE_GRAPH 工具^[12]对模型进行固化压缩,仅保留模型中用于分类的部分参数,缩减模型体积。

(2) 启动微服务预加载以避免多次加载模型。利用 BERT-AS-SERVICE 工具^[13]预加载模型,并启动微服务形式供实时调用。

(3) 优化批量文献处理流程以减少调用次数。进行批量文献分类时,系统首先将第一层分类器的分类结果按大类进行归并,之后统一批量输入到第二层分类器集群中进行分类。例如,将所有通过根分类器分类为 O 大类的文献归并在一起,再将这些文献投入到 O 大类的领域分类器中进行第二次分类,以进一步提升分类速度。

笔者之前已在相关研究中针对医学领域的部分文献数据进行了分类速度测试^[5],此次分别对单篇文献和多篇文献进行分类速度测试,并将微服务与批处理方式和 BERT 原始预测方式进行对比分析,结果如表 4 所示。对于单篇文献分类速度的测量,使用微服务方式可以显著提升分类速度,分类时间从 65 秒左右降至 0.5 秒左右。选取 600 篇文献进行批量文献的分类速度测量,微服务方式可将分类时间由 114 分钟大幅缩减至 8 分钟,采用基于微服务的处理流程优化方式,可进一步将分类时间缩短至 32.9 秒左右。

表 4 文献分类平均时间对比

	BERT 原始预测方式	微服务方式	微服务+处理流程优化
单篇文献	65.195 4 秒	0.514 3 秒	0.514 3 秒
多篇文献	114 分钟	8 分钟	32.911 5 秒

2.4 基于开放接口支撑引擎的开放调用

为了使科技文献自动分类引擎能够得到更加便捷、广泛的使用,笔者设计了对外开放接口。基于 HTTP 协议,对科技文献分类引擎提供 GET 和 POST 两种方式的中文科技文献分类 API 接口服务。API 访问的 URL 统一为 [http://](http://sciengine.las.ac.cn/paper_classification_cn)

sciengine.las.ac.cn/paper_classification_cn,请求参数和返回数据统一为 JSON 格式。

对于 GET 方法,用户以 JSON 格式上传单篇科技文献摘要文本,系统以 JSON 格式返回该文本的分类结果。详细参数信息及示例如表 5 所示。

表 5 GET API 参数信息

参数类型	名称	类型	描述	示 例
请求参数	data	string	文献摘要文本	<code>{"data": "在农业保险中,苹果园受灾理赔需要通过快速准确的落果计数进行定损..."}</code>
返回参数	results	string	文献分类结果	<code>{"results": ["S661.1 苹果", "S436.6 果树病虫害"]}</code>

对于 POST 方法,其 API 参数与 GET 方法的参数保持一致,用户需以列表(list)形式批量上传多篇科技文献摘要文本,系统以 JSON 格式返回每一篇科技文献摘要对应的分类结果。

3 系统实际实现及效果测评

基于科技文献自动分类引擎构建的思路和方法,本文设计和实现了面向实际应用的科技文献中图法自动分类引擎。目前,该分类引擎已在科技文献知识人工智能引擎网站^[14,15]正式推出。具体的系统实现及效果测评如下。

3.1 系统实际实现

本文所构建的面向实际应用的科技文献分类引擎已投入使用并提供对外服务,访问网址为 http://sciengine.las.ac.cn/Cla_2_Layer。网站上提供了中文科技文献分类的在线分类系统和客户端下载服务。

在线分类系统能够对输入的单篇中文科技文献摘要进行分类,通过调用两层分类器集群实时输出分类预测结果。在微服务体系结构的支持下,此响应通常在 0.5 秒内即可完成。

科技文献分类引擎客户端^[16]致力于为科研用户提供便捷易用的操作方式,为科技文献准确、自动分配中图法分类号。客户端根据平台

的 API 接口,提供相应的科技文献分类服务,实现批量科技文献分类的全自动化过程。

3.2 实际分类效果测评

为了验证科技文献分类引擎的实际应用效果,笔者分别在期刊论文分类与学位论文分类上进行分类效果测评。

3.2.1 期刊论文分类的测评及结果

选取另外的与训练数据无交叉的 3 200 篇医学领域的期刊文献进行实际的分类效果测试,将经科技文献分类引擎预测的中图法分类号与其原作者给出的中图法分类号进行对比,统计分类的一致情况。

如表 6 所示,在选取的 3 200 篇医学期刊论文中,科技文献分类引擎的分类结果与作者所给分类号完全一致的论文有 2 277 篇,与作者所给分类号在大类上一致、小类上不一致的论文有 890 篇,与作者所给分类号大类不一致的论文有 33 篇,仅占 1.031%。

对分类不一致的情况进行分析发现,绝大多数经科技文献分类引擎分类与原始分类号有差异的文献,其文摘内容与预测的分类号之间并非毫无关联。例如,某篇文献的摘要为:

聚乙二醇(PEG)修饰是解决蛋白质药物溶解度低、稳定性差、半衰期短、具有免疫原性等问题有效方法,通常在氨基上进

表 6 3 200 篇期刊论文测评结果

结果类型	分类正确数量(篇)	百分比(%)
与作者所给分类号完全一致	2 277	71.156
与作者所给分类号在大类上一致,小类上不一致	890	27.813
与作者所给分类号在大类上不一致	33	1.031

行修饰,但在巯基上进行定点修饰更有利于获得结构明确、组成稳定的修饰产物。综述了聚乙二醇定点修饰蛋白质药物中巯基,包括在游离巯基、二硫键和引入的巯基上进行修饰的研究进展。

作者给出的原始分类号为“R914 药物化学”,而科技文献分类引擎预测的分类号为“Q51 蛋白质,TQ46 制药化学工业”,摘要中包含了蛋白质、药物化学等相关信息,因此预测的分类号具有一定的可解释性。

3.2.2 学位论文分类的测评及结果

为了进一步验证科技文献分类引擎的实际

应用效果,选取计算机科学领域的2473篇学位论文进行实际的分类效果测试。由于学位论文并没有原本标注好的中图法分类号,因此通过随机抽样选取学位论文进行人工审核,判别科技文献分类引擎对其分类是否正确,并统计分类一致的数量。

分类结果有一个分类号和两个分类号的情况,在这两种情况中各随机抽取50篇进行人工判别,具体情况如表7所示。其中,仅有1篇论文与人工判别结果不符,但引擎也将其分到了计算机技术(TP)类目下,说明该引擎仍具有一定的健壮性。

表 7 100 篇人工抽样学位论文的测评结果

人工判别结果	引擎分类结果	
	结果为一个分类号	结果为两个分类号
与人工判别结果完全一致	41	35
与人工判别结果不完全一致(结果正确,但应该有两个类目)	8	—
与人工判别结果不完全一致(两个结果中只有一个相同)	—	15
与人工判别结果不符	1	0

从期刊论文和学位论文的综合测评效果来看,该科技文献分类引擎的分类精度已初步达到实际应用的水平。尤其是在学位论文数据库没有标注中图法分类号的情况下,利用本文所构建的科技文献分类引擎为学位论文自动标注中图法分类号更具有现实意义。

4 结语

在中图法纷繁复杂的类目体系下,自动为

科技文献准确、快速地分配分类号,可以节省长期以来手工标注的成本,进而实现科技文献分类的全自动化过程。本文提出的基于多层分类器集群的科技文献自动分类引擎,能够针对科技文献实现两千多个类目的高效精准分类。引擎还可以根据模型预测得分,为每篇文献最多推荐四个分类号,在单标签分类的基础上进一步体现了学科交叉特性。

本文重点对科技文献自动分类引擎的整体架构、引擎建设的四个关键技术问题及解决方

案、系统实现及效果测评进行了阐述。基于层次分类思想,设计和开发了面向实际应用的科技文献中图法自动分类引擎,该引擎能够将文献自动分类到中图法三、四级类目下,且通过实际测试的检验,各项指标基本达到了实用要求,具有实际的应用价值。

当然,基于这一思路开发的面向实际应用

的科技文献中图法自动分类引擎还有需要改进和提高之处。如在学科领域覆盖方面,可以通过进一步收集语料资源覆盖更多的类目;在科技文献语种方面,可以考虑利用英文科技文献摘要数据进行分类模型的训练。这也是未来研究的重点所在。

参考文献

- [1] 冉亚鑫,韩红旗,张运良,等. 基于 Stacking 集成学习的大规模文本层次分类方法[J]. 情报理论与实践, 2020,43(10):171-176,182. (Ran Y X, Han H Q, Zhang Y L, et al. Large scale text hierarchical classification method based on Stacking ensemble learning[J]. Information Studies: Theory & Application, 2020,43(10): 171-176,182.)
- [2] 王昊,严明,苏新宁. 基于机器学习的中文书目自动分类研究[J]. 中国图书馆学报,2010,36(6):28-39. (Wang H, Yan M, Su X N. Research on automatic classification for Chinese bibliography based on machine learning[J]. Journal of Library Science in China, 2010,36(6):28-39.)
- [3] 王昊,叶鹏,邓三鸿. 机器学习在中文期刊论文自动分类研究中的应用[J]. 现代图书情报技术,2014(3): 80-87. (Wang H, Ye P, Deng S H. The application of machine-learning in the research on automatic categorization of Chinese periodical articles[J]. Data Analysis and Knowledge Discovery, 2014(3):80-87.)
- [4] 刘高军,陈强强. 基于极限学习机和混合特征的中文书目自动分类模型研究[J]. 北方工业大学学报, 2018,30(5):99-104. (Liu G J, Chen Q Q. Research on automatic classification model of Chinese bibliography based on extreme learning machine and mixed features[J]. Journal of North China University of Technology, 2018,30(5):99-104.)
- [5] 赵旻,张智雄,刘欢. 基于层次分类法的中文医学文献分类研究[J]. 图书馆学研究,2021(21):49-55,61. (Zhao Y, Zhang Z X, Liu H. Research on the classification of Chinese medical literature based on hierarchical classification[J]. Research on Library Science, 2021(21):49-55,61.)
- [6] Devlin J, Chang M W, 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 (NAACL-HLT): Human Language Technologies. Minneapolis, Minnesota, 2019:4171-4186.
- [7] Shi K Z, Gong C J, Lu H, et al. Wide-grained capsule network with sentence-level feature to detect meteorological event in social network[J]. Future Generation Computer Systems, 2020(102):323-332.
- [8] Chen F, Yuan Z, Huang Y. Multi-source data fusion for aspect-level sentiment classification[J]. Knowledge-Based Systems, 2020(187):104831.
- [9] Zhang X, Zhang Y, Zhang Q, et al. Extracting comprehensive clinical information for breast cancer using deep learning methods[J]. International Journal of Medical Informatics, 2019(132):103985.
- [10] Du Y P, Pei B B, Zhao X Z, et al. Deep scaled dot-product attention based domain adaptation model for biomedic-

- al question answering[J]. Methods, 2020(173):69-74.
- [11] Moradi M, Dorffner G, Samwald M. Deep contextualized embeddings for quantifying the informative content in biomedical text summarization[J]. Computer Methods and Programs in Biomedicine, 2020(184):105117.
- [12] TensorFlow[EB/OL]. [2021-09-03]. https://github.com/tensorflow/tensorflow/blob/master/tensorflow/python/tools/freeze_graph.py.
- [13] hanxiao. Bert-as-service[EB/OL]. [2021-09-03]. <https://github.com/hanxiao/bert-as-service>.
- [14] 张智雄, 刘欢, 于改红. 构建基于科技文献知识的人工智能引擎[J]. 农业图书情报学报, 2021, 33(1):17-31. (Zhang Z X, Liu H, Yu G H. Building an artificial intelligence engine based on scientific and technological literature knowledge[J]. Journal of Library and Information Science in Agriculture, 2021, 33(1):17-31.)
- [15] 科技文献知识人工智能引擎[EB/OL]. [2021-09-03]. <http://sciengine.las.ac.cn>. (Scientific and technological literature knowledge artificial intelligence engine[EB/OL]. [2021-09-03]. <http://sciengine.las.ac.cn>.)
- [16] 科技文献人工智能引擎客户端[EB/OL]. [2021-09-03]. <http://sciengine.las.ac.cn/download/client.zip>. (Scientific and technological literature knowledge artificial intelligence engine client[EB/OL]. [2021-09-03]. <http://sciengine.las.ac.cn/download/client.zip>.)

张智雄 中国科学院文献情报中心副主任、研究馆员, 中国科学院大学经济与管理学院图书情报与档案管理系教授, 博士生导师。北京 100190。

赵 旻 中国科学院文献情报中心、中国科学院大学经济与管理学院图书情报与档案管理系博士研究生。北京 100190。

刘 欢 中国科学院文献情报中心、中国科学院大学经济与管理学院图书情报与档案管理系博士研究生。北京 100190。

(收稿日期:2021-09-28;修回日期:2022-05-06)