doi:10.3772/j.issn.2095-915x.2021.04.010

大规模文本分类的训练语料去噪方法研究

高雄[1,2] 韩红旗[1,2] 王力[1,2] 薛陕[1,2]

1. 中国科学技术信息研究所 北京 100038；
2. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038

开放科学（资源服务）标识码（OSID）

摘要： [目的 / 意义] 训练语料的质量对主流的文本分类算法至关重要。消除噪声，尤其是“类别外噪声”，有助于提升训练语料的质量，进而提升文本分类算法的准确率。[方法 / 过程] 本文重点利用语义信息来消除“类别外噪声”。通过对每个类别的训练语料构建“类目 - 类目关键词”知识库，利用“词嵌入”自动化比较其语义信息来判断该类别下是否存在噪声，并给出“类别外噪声”类目候选列表以及文献候选列表，最后通过人机交互的方式消除噪声。[结果 / 结论] 本文提出的去噪方法能够有效检测并消除大规模文本分类的训练语料中的噪声数据，提升训练语料的质量。

关键词： 文本分类；去噪；词嵌入

中图分类号： TP391，G35

Research on Denoising Method of Training Corpus for Large-scale Text Classification

GAO Xiong[1,2] HAN Hongqi[1,2] WANG Li[1,2] XUE Shan[1,2]

1. Institute of Scientific and Technical Information of China, Beijing 100038, China;
2. Key Laboratory of Rich-media Knowledge Organization & Service of Digital Publishing Content, Beijing 100038, China

基金项目 中国科学技术信息研究所创新研究基金青年项目“大规模文本分类的训练语料去噪研究”（QN2020-10）；中国工程科技知识中心建设项目“知识组织体系建设”（CKCEST-2021-2-6）。

作者简介 高雄（1990-），硕士，助理研究员，研究方向为文本挖掘，E-mail：gaox@istic.ac.cn；韩红旗（1971-），博士，研究员，研究方向为知识组织与知识工程；王力（1993-），硕士，助理研究员，研究方向为知识服务；薛陕（1984-），博士，助理研究员，研究方向为知识组织与知识工程、自然语言处理、文本数据挖掘。

引用格式 高雄，韩红旗，王力，等 . 大规模文本分类的训练语料去噪方法研究 [J]. 情报工程，2021, 7(4): 117-126.

doi:10.3772/j.issn.2095-915x.2021.04.010

Abstract: [Objective/ Significance] The quality of the training corpus is very important to mainstream text classification algorithms. Denoising, especially "out-of-category noise", helps to improve the quality of the training corpus, thereby improving the accuracy of the text classification algorithm. [Methods/Process] This paper focuses on using semantic information to eliminate "out-of-category noise", builds a "category-category keywords" knowledge base for each category of training corpus, and uses "word embedding" to automatically compare its semantic information to determine whether there is noise in the category. And give a list of candidates for noise category and noise data, and finally eliminate the noise by means of human-computer interaction. [Results /Conclusions] The denoising method proposed in this paper can effectively detect and eliminate the noise data in the training corpus of large-scale text classification, and improve the quality of the training corpus.
Keywords: Text classification; denoising;word embedding

引言

随着互联网数据规模的不断增长，大规模文本分类技术已成为当今大数据时代迫切需要的关键技术之一，并引起了学者的广泛关注[1,2]。普通的文本分类问题一般只涉及数个类别，多则数十个类别。与普通文本分类问题不同，大规模文本分类的类别数量较多，如《中国图书馆分类法》（原称《中国图书馆图书分类法》，以下简称《中图法》）包括的类目数量达到两万以上，其中的三级类目接近 2500 个，使用一般的方法很难对数千个甚至上万个类别进行准确地分类[3]。因此如何选择一种有效的方法对大规模文本进行分类以达到理想的效果是目前文本分类领域亟待解决的关键问题。最近几年，基于神经网络的分类算法在文本分类领域得到应用并取得不错的效果[4-7]。然而，这种学习算法需要每个类别下具有大量的训练样本数据。也就是说，目前主流的文本分类算法的效果很大程度上取决于训练语料的质量。从某种意义上讲，需要的训练语料越多，其所包含噪声的概率也就越大。因此，在将训练语料输入深度学习模型之前，对训练语料进行清洗是非常重要的，这直接决定了模型的效果，甚至会影响模型的收敛。

数据清洗，也称为去噪，就是把冗杂、混乱、无效的噪声数据清洗干净。郭宇航[8]使用基于语言模型和基于点互信息等两种方式消除基于同义词扩展的训练语料中的噪声。Xu[9]等人于 2019 年在认真分析了 Web of Science（WoS）中的被引文献的 DOI 的错误类型及其发生的频率的基础上，提出了一种基于正则表达式的数据清洗的方法，实验结果表明该方法能够有效提高训练语料的质量。

目前专门针对大规模文本分类中的训练语料去噪问题的研究还较少，多集中于简单的基于正则表达式的方法，即通过分析常见的噪声类型，人工编写正则表达式规则，从而去除训练数据中的噪声；或是通过语言模型等方法，即计算一个词语序列构成一个句子的概率，来判断一句话出现的概率高不高，是否符合日常的表达习惯等。这些方法只能去除"错别字、乱码、语法错误和不符合所属语言或领域的表达习惯"带来的噪声，而大规模文本分类中的训练语料的噪声多是"类别外噪声"，如：中图分类号 A14（一级类目 A：马克思主义、列

宁主义、毛泽东思想、邓小平理论；二级类目A1：马克思、恩格斯著作；三级类目A14：诗词）类别的训练语料中出现的“基于‘地区’特点的风险投资环境的系统分析研究”这样的“语义”上明显不属于A14的数据。显然，“类别外噪声”是无法通过基于正则表达式或是语言模型的方法来消除的。

本文重点利用语义信息来消除“类别外噪声”，通过对每个类别的训练语料构建“类目-类目关键词”知识库，利用“词嵌入”自动化比较其语义信息来判断该类别下是否存在噪声，并给出噪声类目候选列表和噪声数据候选列表，最后通过人机交互的方式消除噪声，从而提升大规模文本分类中的训练语料质量，进而不断提高自动分类标引工具的准确率，更好地在科技领域服务于科研人员。

1 方法设计

本文基于《中图法》对大规模文本分类的训练语料进行去噪研究。分为“构建分类文件系统”、“分析噪声类型与多策略去噪”等部分，其中重点研究“类别外噪声”的消除。

1.1 构建分类文件系统

在实践中，大规模文本分类的训练语料多为不同类别（如不同的中图分类号）的文献混杂在一起，存储在纯文本文件或SQL文件中，这样不利于观察训练语料中的噪声类型及其特征，从而影响去噪。因此，有必要按照《中图法》将各个类别混在一起的训练语料分门别类地放在不同的文件夹以及文件中，即按照《中图法》构建分类文件系统。

《中图法》是我国建国后编制出版的一部具有代表性的大型综合性分类法，是当今国内图书馆使用最广泛的分类法体系。《中图法》出版于1975年，2010年出版了第五版。《中图法》共分5个基本部类、22个大类，采用英文字母与阿拉伯数字相结合的混合号码（即中图分类号），用一个字母代表一个大类，以字母顺序反映大类的次序，在字母后用数字作标记。

根据《中图法》的分类结构以及文本分类的精度要求，本文建立层级粒度到三级中图分类号的分类文件系统，因此首先需要构建三级类目嵌套字典。其结构如下：

1）三级字典（最内层字典）

键：三级中图分类号（形如：A11），值：1；

2）二级字典

键：二级中图分类号（形如：A1），值：三级字典；

3）一级字典（最外层字典）

键：一级中图分类号（形如：A），值：二级字典。

需要特别说明的是，如G51（世界教育事业，三级中图分类号）的上级中图分类号并不是G5，这样不利于根据三级类目嵌套字典对训练语料进行抽取。为了解决这一问题，增加了10个'*_NA'二级类目号，10个'*_NA'二级类目号及其对应的三级类目号具体如下：

1）G5_NA：G51/57、G51、G52、G53/57

2）G6_NA：G61/79、G61、G62、G63、G64、G65

3）G7_NA：G71、G72、G74、G75、G76、G77、G78、G79

doi:10.3772/j.issn.2095-915x.2021.04.010

4）H82_NA：H824

5）K8_NA：K82、K833/837、K86、K87/879.49、K87、K883/887

6）O2_NA：O21、O22、O23、O24、O29

7）O5_NA：O51、O52、O53、O55、O56、O57、O59

8）P6_NA：P61、P62、P64、P68、P691、P694

9）V3_NA：V31、V32、V35、V37

10）V5_NA：V51、V52、V55、V57

根据上述构建的三级类目嵌套字典对训练语料进行抽取，构建基于《中图法》的分类文件系统。具体步骤如下：获取训练语料的必要字段（如：id、标题、关键词等）以及中图分类号列表；依次循环上述构建的三级类目嵌套字典的一、二、三级类目号，并判断“训练语料中的记录是否属于该级类目号”，若属于该级类目号，则划分到以该级类目号命名的 txt 文件中。构建的分类文件系统结构如下：

1）一级类目号对应文件夹；

2）二级类目号对应文件夹，存放于该二级类目号属于的一级类目号文件夹下；

3）三级类目号对应 txt 文件，存放于该三级类目号属于的二级类目号文件夹下。

为了便于统计，文件夹名或者 txt 文件名均以“_”+ 该类目下的记录数结尾。

1.2 分析噪声类型与多策略去噪

（1）分析噪声类型

基于上述构建好的“分类文件系统”，随机抽取一些训练语料文件进行观察，分析并总结噪声类型如表 1 所示。其中除了包括常见的训练语料中的噪声类型：出现字母、html 标签、多余的空格和特殊符号等，还有“类别外噪声”，如：中图分类号 A14（一级中图分类号 A：马克思主义、列宁主义、毛泽东思想、邓小平理论；二级中图分类号 A1：马克思、恩格斯著作；三级中图分类号 A14：诗词）类别的训练语料中出现的“基于‘地区’特点的风险投资环境的系统分析研究”这样的“语义”上明显不属于 A14 的数据。

表 1 噪声定义及编号

噪声编号	噪声定义
1	出现字母
2	出现html标签
3	出现多余的空格
4	出现特殊符号（连接符、下划线、英文问号等）
5	出现特殊符号不配对以及半角全角不统一的情况
6	类别外噪声

（2）多策略去噪

针对（1）部分分析出的不同噪声类型，本文采用不同的策略针对性地消除不同类型的噪声。采用基于正则表达式的方法，可以相对容易地将上述编号 1-5 类型的噪声消除，而针对“类别外噪声”，则无法通过基于正则表达式的方法消除，该类型噪声的消除也是本文的重点。下文重点阐述该类型噪声的消除。

1.3 消除“类别外噪声”

（1）构建“类目 - 类目关键词”知识库

将《中图法》中每个中图分类号的语义描述按照层级关系排列与对应的三级中图分类号（一个三级中图分类号即是一个类目）组成一

个字段，即“类目语义描述”字段。基于上述构建好的分类文件系统，对每个三级中图分类号对应的训练语料中的关键词字段做词频统计，最终返回最高频的前 5 个关键词作为类目关键词，是为“类目 - 类目关键词”知识库的第 2 个字段，即“类目关键词”字段。下文使用类目关键词作为识别“类别外噪声”的特征，如果类目关键词和对应的三级中图分类号的“类目描述”在语义上不相关，则表明该类目对应的训练语料中存在大量的噪声数据。

（2）两阶段法消除“类别外噪声”

针对大规模文本分类的训练语料的“类别外噪声”，本文采用“两阶段法”逐步定位“类别外噪声”。第 1 阶段：给出可能存在“类别外噪声”的类目候选列表。基于上述构建的“类目 - 类目关键词”知识库，使用“类目关键词中的字符没有出现在类目语义描述”的约束，对知识库中的每一个类目进行初筛。对于初筛之后的“（类目语义描述，类目关键词）字符串对”使用 Bert 预训练模型 [10] 进行“词嵌入”，得到其向量化表示，计算两者之间的余弦相似度并根据该值对“字符串对”进行排序，从而自动化比较“类目描述”的语义以及其对应的“类目关键词”的语义是否相似，来判断该类别下是否存在“类别外噪声”，然后将可能存在“类别外噪声”的类目候选列表交由人工审核确定阈值，相似度值小于该阈值的需要重点审核；第 2 阶段：给出可能存在“类别外噪声”的文献候选列表。根据审核后的存在“类别外噪声”的类目列表以及 1.1 部分构建好的分类文件系统，定位到存在“类别外噪声”的类目训练语料文件，对于十分明显的“类别外噪声”文献的训练语料文件，可以通过关键词匹配的方式直接消除；对于混杂“类别外噪声”文献的训练语料文件，使用 Bert 预训练模型对该文件中的不同的文献的关键词进行“词嵌入”得到其向量，使用余弦相似度来计算文献之间的距离，给出可能存在“类别外噪声”的文献候选列表，最后通过人机交互的方式消除噪声。

2 实验及结果分析

2.1 实验数据集及设置

本文实验所用数据为收集的约 800 万条覆盖 22 个一级中图分类号的会议论文、期刊论文等文献数据，包含题名、关键词、摘要、中图分类号等信息。本文所用的数据中，一级中图分类号 22 个，二级中图分类号 254 个，三级中图分类号 1679 个。数据集关于一级中图分类号的统计信息如表 2 所示。

表 2 数据集的统计信息

中图分类号（一级中图分类号）	论文数量	中图分类号（一级中图分类号）	论文数量
A	9202	N	12298
B	37943	O	404522
C	47505	P	218307
D	112407	Q	232447
E	12223	R	2907509
F	413146	S	453525
G	294342	T	2409782
H	30120	U	130404
I	38993	V	66654
J	23860	X	184957
K	33927	Z	278

doi:10.3772/j.issn.2095-915x.2021.04.010

在“两阶段法消除‘类别外噪声’”部分使用的是主流的 Google 公布的一个具有 12 层，768 维，1.1 亿参数的中文 BERT 预训练模型 chinese_L-12_H-768_A-12。

2.2 实验结果及分析

针对实验数据集构建的层级粒度到三级中图分类号的分类文件系统，以 A（一级中图分类号）、A1（二级中图分类号）、A11（三级中图分类号）为例，如图 1 所示，其中的文件夹名或者 txt 文件名均以“_”+ 该类目下的记录数结尾。根据分类文件系统可以清楚地看出实验数据集中一、二、三级中图分类号对应的文献数量。

图 1 实验数据集构建的分类文件系统结构示意图

针对实验数据集构建的“类目 - 类目关键词”知识库，以 A14 举例为：“类目语义描述”字段：A14（马克思主义、列宁主义、毛泽东思想、邓小平理论 / 马克思、恩格斯著作 / 诗词）；“类目关键词”字段：风险投资、环境、系统分析、地区、速效磷。

（1）消噪第 1 阶段实验

在两阶段法消除“类别外噪声”的第 1 阶段中，经过使用“类目关键词中的字符没有出现在类目语义描述”的约束初筛之后，1679 个三级中图分类号（类目编号：1-1679）中有 234 个可能存在“类别外噪声”的类目，如表 3 所示。使用 Bert 预训练模型向量化“类目语义描述”与“类目关键词”，并计算两者之间的余弦相似度并根据该值由小到大排序后的结果，如表 4 所示。经由专家审核确定的阈值为 0.82（对应表 4 中的序号为 90 的类目），即可能存在“类别外噪声”的类目候选列表中余弦相似度值高于 0.82 的 146 个类目，除极个别（11 个类目）外，均不存在“类别外噪声”，需要重点审核相似度值小于该阈值的 89 个类目。经审核，234 个类目中确定存在“类别外噪声”的类目数量为 44 个（见表 5）。

（2）消噪第 2 阶段实验

第 2 阶段：给出可能存在“类别外噪声”的文献候选列表。根据人工审核后的存在“类别外噪声”的类目列表，定位到对应的类目训练语料文件，对于十分明显的“类别外噪声”文献的训练语料文件，可以通过关键词匹配的方式直接消除，如类目 A14（马克思主义、列宁主义、毛泽东思想、邓小平理论 / 马克思、恩格斯著作 / 诗词），一共 4 条文献记录（见表 6），都与 A14 语义不符，可直接消除。

表 **3**　初筛之后可能存在“类别外噪声”的类目候选列表

序号	类目编号	类目语义描述	类目关键词
1	4	A14（马克思主义、列宁主义、毛泽东思想、邓小平理论/马克思、恩格斯著作/诗词）	风险投资　环境　系统分析　地区　速效磷
2	5	A15（马克思主义、列宁主义、毛泽东思想、邓小平理论/马克思、恩格斯著作/手迹）	细菌　土壤胶体　矿物体系　铜(Cu2+)　吸附
3	11	A25（马克思主义、列宁主义、毛泽东思想、邓小平理论/列宁著作/手迹）	单端匹配　功率　微波单片集成电路　单刀单掷　PIN二极管
4	13	A28（马克思主义、列宁主义、毛泽东思想、邓小平理论/列宁著作/语录）	斑蝥　药性　中药　岗松　薄层色谱
5	34	A58（马克思主义、列宁主义、毛泽东思想、邓小平理论/马克思、恩格斯、列宁、斯大林、毛泽东、邓小平著作汇编/语录）	利肺片　抗结核药物　治疗　糖尿病合并肺结核
6	38	A74（马克思主义、列宁主义、毛泽东思想、邓小平理论/马克思、恩格斯、列宁、斯大林、毛泽东、邓小平生平和传记/斯大林）	抑郁症　反补贴　非市场经济国家　中美反补贴案　急性ST段抬高型心肌梗死(STEMI)
7	53	B12（哲学、宗教/世界哲学/古代哲学）	知识　柏拉图　苏格拉底　辩证法　道
8	55	B14（哲学、宗教/世界哲学/近代哲学）	自然　笛卡尔　启蒙　康德　认识论
9	61	B22（哲学、宗教/中国哲学/先秦哲学（~前220年））	孔子　儒家　老子　道　庄子
10	73	B302（哲学、宗教/亚洲哲学/古代哲学）	俄罗斯经济　中俄关系　俄罗斯东部开发　东北振兴　中俄合作
…	…	…	…
234	1664	Z42（综合性图书/论文集、全集、选集、杂著/中国论文集、全集、选集、杂著）	设计　项目经理　管理　P2P　Chord

表 **4**　排序之后可能存在“类别外噪声”的类目候选列表

序号	类目编号	类目语义描述	类目关键词	余弦相似度值
1	11	A25（马克思主义、列宁主义、毛泽东思想、邓小平理论/列宁著作/手迹）	单端匹配　功率　微波单片集成电路　单刀单掷　PIN二极管	0.6444502
2	34	A58（马克思主义、列宁主义、毛泽东思想、邓小平理论/马克思、恩格斯、列宁、斯大林、毛泽东、邓小平著作汇编/语录）	利肺片　抗结核药物　治疗　糖尿病合并肺结核	0.659111
3	13	A28（马克思主义、列宁主义、毛泽东思想、邓小平理论/列宁著作/语录）	斑蝥　药性　中药　岗松　薄层色谱	0.66832393
4	5	A15（马克思主义、列宁主义、毛泽东思想、邓小平理论/马克思、恩格斯著作/手迹）	细菌　土壤胶体　矿物体系　铜(Cu2+)　吸附	0.6732846
5	4	A14（马克思主义、列宁主义、毛泽东思想、邓小平理论/马克思、恩格斯著作/诗词）	风险投资　环境　系统分析　地区　速效磷	0.67883074
6	725	K23（历史、地理/中国史/封建社会（公元前475~公元1840年））	汉代　北魏　魏晋南北朝　西汉　秦汉	0.70790076
7	38	A74（马克思主义、列宁主义、毛泽东思想、邓小平理论/马克思、恩格斯、列宁、斯大林、毛泽东、邓小平生平和传记/斯大林）	抑郁症　反补贴　非市场经济国家　中美反补贴案　急性ST段抬高型心肌梗死(STEMI)	0.7232864
8	61	B14（哲学、宗教/世界哲学/近代哲学）	孔子　儒家　老子　道　庄子	0.73211634
9	715	K11（历史、地理/世界史/上古史（公元前40世纪以前））	信用卡　持卡人　发卡机构　刷卡套现　盗刷	0.7568302
10	741	K37（历史、地理/亚洲史/西亚（西南亚））	伊朗　土耳其　塞浦路斯　阿富汗　伊斯兰革命	0.75838536
…	…	…	…	…
90	1434	TL2（工业技术/原子能技术/核燃料及其生产）	铀　地浸采铀　铀矿石　堆浸　地浸	0.82210803
…	…	…	…	…
234	717	K13（历史、地理/世界史/中世纪史（476~1640年））	中世纪　宗教改革　西欧　基督教　赋税基本理论	0.9045619

doi:10.3772/j.issn.2095-915x.2021.04.010

表 5　人工审核后的存在“类别外噪声”的类目列表

序号	类目编号	类目语义描述	类目关键词	余弦相似度值
1	11	A25（马克思主义、列宁主义、毛泽东思想、邓小平理论/列宁著作/手迹）	单端匹配　功率　微波单片集成电路　单刀单掷　PIN二极管	0.6444502
2	34	A58（马克思主义、列宁主义、毛泽东思想、邓小平理论/马克思、恩格斯、列宁、斯大林、毛泽东、邓小平著作汇编/语录）	利肺片　抗结核药物　治疗　糖尿病合并肺结核	0.659111
3	13	A28（马克思主义、列宁主义、毛泽东思想、邓小平理论/列宁著作/语录）	斑蝥　药性　中药　岗松　薄层色谱	0.66832393
4	5	A15（马克思主义、列宁主义、毛泽东思想、邓小平理论/马克思、恩格斯著作/手迹）	细菌　土壤胶体　矿物体系　铜(Cu2+)　吸附	0.6732846
5	4	A14（马克思主义、列宁主义、毛泽东思想、邓小平理论/马克思、恩格斯著作/诗词）	风险投资　环境　系统分析　地区　速效磷	0.67883074
6	38	A74（马克思主义、列宁主义、毛泽东思想、邓小平理论/马克思、恩格斯、列宁、斯大林、毛泽东、邓小平生平和传记/斯大林）	抑郁症　反补贴　非市场经济国家　中美反补贴案　急性ST段抬高型心肌梗死(STEMI)	0.7232864
7	715	K11（历史、地理/世界史/上古史（公元前40世纪以前））	信用卡　持卡人　发卡机构　刷卡套现　盗刷	0.7568302
8	633	I239（文学/中国文学/曲艺）	政府审计　公共治理　有效政府　有限政府　透明政府	0.7638516
9	594	H78（语言、文字/印欧语系/凯尔特语族）	干旱胁迫　硅　灌水量　蒸散量	0.76948416
10	773	K73（历史、地理/美洲史/拉丁美洲）	肺肿瘤　诊断　免疫组织化学　肝癌　腹腔镜	0.7769891
…	…	…	…	…
44	827	N45（自然科学总论/自然科学教育与普及/教学实验、实习）	农民起义　气候冲击　政府能力　空间位置记忆广度　线性回归	0.8417687

表 6　全部为“类别外噪声”的类目训练语料文件（以 **A14** 为例）

文献编号	标题	关键词
1	基于"地区"特点的风险投资环境的系统分析研究	风险投资;环境;系统分析;地区
2	我国农业生产对磷肥的需求现状及展望	速效磷;供磷水平;磷肥;年需求量;发展前景
3	群落结构和叶面积指数在具茨山植被次生演替中的变化	具茨山;生物多样性;叶面积指数;退耕还林;生态恢复;生态服务功能
4	从土壤肥力变化预测中国未来磷肥需求	磷平衡;肥力变化;磷肥需求

对于混杂“类别外噪声”文献的训练语料文件，如类目 H77（语言、文字 / 印欧语系 / 罗马语族），一共 10 条文献记录（见表 7），有些文献是“类别外噪声”（如表 7 中的 1、2、10），而另一些不是，需要使用“词嵌入”并计算余弦相似度，给出可能存在“类别外噪声”的文献候选列表（见表 8），经专家审核，候选列表中的前 7 条均为“类别外噪声”文献。

（3）结果分析

表 3 中类目编号 4、5、11、13、34、38、73 等 7 个例子表明使用约束可以找出明显是“类别外噪声”的类目，而类目编号 53、55、61 等 3 个例子表明使用约束初筛之后类目候选列表中会存在一些正确无噪声的类目，所以依然有必要使用“词嵌入”、计算相似度等方式对候选列表进行排序。

表 7　混杂“类别外噪声”的类目训练语料文件（以 **H77** 为例）

文献编号	标题	关键词
1	基于ZigBee技术的无导联线动态心电监测仪设计	动态心电监测;无导联线;无线传输;ZigBee
2	利用胶原-肝素自组装多层膜改善纯钛表面血液相容性的研究	钛;层层自组装(LBL);胶原;肝素;血液相容性
3	汉语—意大利语委婉语对比研究	人类学;中国;意大利;委婉语;对比研究
...	...	...
10	微创胸腔镜与传统开胸手术治疗肺癌的疗效对比分析	肺癌;微创胸腔镜;肺癌根治术

表 8　可能存在“类别外噪声”的文献候选列表（以 **H77** 为例）

序号	文献编号	类目语义描述	文献的关键词	余弦相似度值
1	10	H77（语言、文字/印欧语系/罗马语族）	肺癌;微创胸腔镜;肺癌根治术	0.741693
2	8		青光眼,闭角型;光学相干断层扫描;视网膜神经纤维层	0.7613801
3	1		动态心电监测;无导联线;无线传输;ZigBee	0.88167465
...	...	...	...	...
10	6	H77（语言、文字/印欧语系/罗马语族）	意大利;文艺复兴;俗语;拉丁语;语言问题	0.89639866

从表 4 可以看出，虽然经过排序之后类目编号 725、61、741 所表示的正确无噪声的类目仍在表中，但是类如 715 等类目编号进入到前 10 中，即将初筛列表中混杂的噪声类目提升到了更显著的位置，方便后续人工审核及确定阈值。表 8 也能体现出类似的情况。

使用本文方法，待去噪的三级类目数量逐步从最初的 1679 个筛选为 234 个，再从 234 个确定为最终的 44 个。此外，对 44 个存在“类别外噪声”的类目进行统计分析，可以得出 H、K、A 三个一级类目下出现的“类别外噪声”较多。

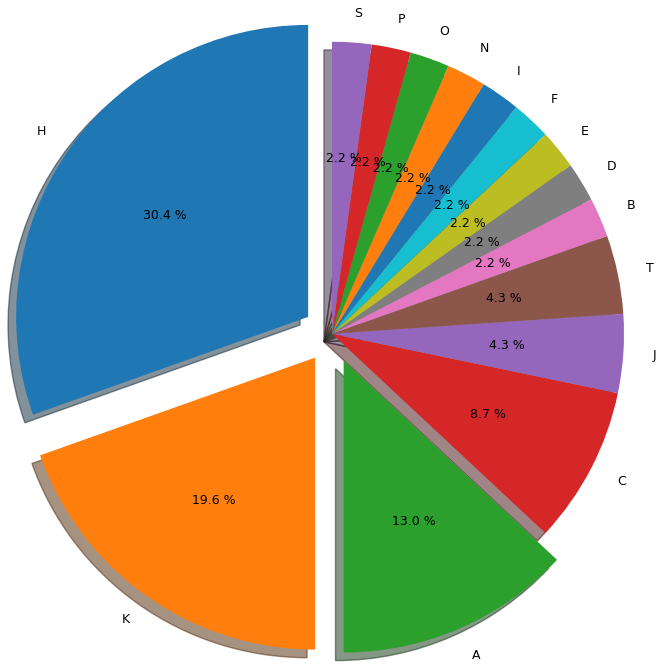

图 2　实验数据集中“类别外噪声”一级类目统计图

doi:10.3772/j.issn.2095-915x.2021.04.010

3 展望

自动分类标引工具的训练目前以监督学习方法为主，而训练语料的质量对于大规模文本分类自动标引工具准确率的提升至关重要。本文重点利用语义信息来消除“类别外噪声”，提出了两阶段法对其进行消除。通过对约 800 万条会议论文、期刊论文等文献数据的去噪实验结果分析，证明此方法可有效检测出存在“类别外噪声”的类目以及文献。大规模文本分类具有类目数量大、类目层次多、数据量大等特点，在此种情境下去噪是一件非常复杂的工作，本研究提出的方法可有效减少人工工作量，快速检测出存在噪声的数据。但因为本方法基于高频关键词，所以对一个具有少量噪声数据的类目的处理则需要新的方法。这种更细致的情况下的研究是未来的工作方向。

参 考 文 献

[1] 李保利 . 基于类别层次结构的多层文本分类样本扩展策略 [J]. 北京大学学报（自然科学版）, 2015, 51(2):357-366.

[2] 刘琼昕 , 宋祥 , 王鹏 . 面向出版社富媒体知识的文本分类研究 [J]. 情报工程 , 2019, 5(2):40-48.

[3] 黄莉 , 李湘东 . 基于《中图法》的自动分类研究现状与展望 [J]. 图书情报知识 , 2012(4):30-36.

[4] 刘婷婷 , 朱文东 , 刘广一 , 等 . 基于深度学习的文本分类研究进展 [J]. 电力信息与通信技术 , 2018, 16(3):1-7.

[5] 殷亚博 , 杨文忠 , 杨慧婷 , 等 . 基于卷积神经网络和 KNN 的短文本分类算法研究 [J]. 计算机工程 , 2018, 44(7):193-198.

[6] 翟文洁 , 闫琰 , 张博文 , 等 . 基于混合深度信念网络的多类文本表示与分类方法 [J]. 情报工程 , 2016, 2(5):30-40.

[7] 代令令 , 蒋侃 . 基于 Fasttext 的中文文本分类 [J]. 计算机与现代化 , 2018(5):35-40,85.

[8] 郭宇航 . 词义消歧语料库自动获取方法研究 [D]. 哈尔滨 : 哈尔滨工业大学 , 2008.

[9] Xu S, Hao L, An X, et al. Types of DOI errors of cited references in Web of Science with a cleaning method[J]. Scientometrics, 2019, 120(3): 1427-1437.

[10] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. In Proceedings of NAACL-HLT. 2019: 4171-4186.