北京郵電大學

Beijing University of Posts and Telecommunications

网络空间安全学院 硕士毕业论文答辩

基于深度学习的交叉学科中频繁项发现与研究

赵秋涵

指导老师：杨文川

论文背景与内容

选题背景

- 本课题来源于事、企业合作项目，项目的初衷旨在帮助北京市某研究所构建来源于知网(CNKI)等文献检索平台的科技文章多标签自动标注功能。

- 随着不同学科的迁移应用与相互融合，产生不少的交叉学科。过去的单一分类方法难以对这些交叉学科中的频繁项进行描述和定位，若帮助人们快速的**定位**和**挖掘**交叉学科中的**频繁项***，将具有极强的应用价值。

频繁项*：指在一批文献中频繁出现的课题或研究方向，而非“频繁项集”概念

论文背景与内容

课题内容

构建合理的科技文献描述符，通过构建深度学习网络获取每篇文献对应的描述符，进一步通过聚类算法对一批文献中的频繁项进行发现与分析。

数据获取与预处理模块

- 数据集获取及预处理 (数据清洗、预训练词向量)

- 科技文献描述符 (多标签+tf-idf关键词)　**深度学习模块**
- 不同多标签神经网络 (TextCNN/RCNN/attention/BERT)

- 聚类算法 (自适应参数DBSCAN)　**聚类模块**

课题相关先行研究

论文成果

[1] Sequence Generative Adversarial Network for Chinese Social Media Text Summarization[A]. 2019 Chinese Automation Congress (CAC)[C]. 2019: 4620–4625.

[2] A Method for Massive Scientific Literature Clustering Based on Hadoop[A]. 2019 Chinese Automation Congress (CAC)[C]. 2019: 5518–5523.

[3] Design and Implementation of Application Classification Based on Deep Learning[A]. 2019 Chinese Automation Congress (CAC)[C]. 2019: 4821–4826.

[4] Design and Research of Composite Web Page Classification Network Based on Deep Learning[A]. 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)[C]. 2019: 1531–1535.

[5] Multi-label classification of technical articles based on deep neural network[A]. 2019 Chinese Control Conference (CCC)[C]. Guangzhou, China: IEEE, 2019: 8391–8397.

课题相关先行研究

项目成果

学科领域新知识发现与分类 (18CNIC-025701-004-06)

前沿科技文献评分推荐 (0686-1941B1521095Z/06)

相关竞赛

科大讯飞2019iFLYTEK-大数据应用分类标注挑战赛 (4th)

KAGGLE Jigsaw Unintended Bias in Toxicity Classification top9 (281th,铜奖)

论文创新点

- **针对中文科技文献的多标签分类网络。**通过引入预训练网络和注意力机制等方式，训练深度学习网络构建科技文章与多标签之间的映射关系。在此之上，本课题通过统计分布的方式，进一步确定多标签预测概率向量的阈值，让每篇科技文献能够对应相应的1-4个知网标签。

- **新的文本特征表示。**本课题采用“多标签加全文关键词”的方式作为文本特征表示。科技论文的“标题和摘要”是含括全文信息的重要内容，而多标签正是对“标题加摘要”这种短文本的高级抽象；而关键词则是对多标签的进一步补充，防止标签过少或过于集中所产生的的信息损失。与之前的工作相比，这种方法产生特征的速度快，且特征涵盖原文的更多的信息。

- **聚类算法的优化。**针对本课题的文本表征，该特征向量的维度较高，难以通过超参数调优的方式确定聚类算法的参数对。基于这一难点，本课题提出了一种适用于所爬取科技文献数据的自适应参数的DBSCAN密度聚类算法。该算法通过核密度估计和基于轮廓系数的优化，能自动的产生较好的DBSCAN参数对。

整体框架

爬取数据集的统计特征

- 分布统计
- 词长度统计

数据增强

深度学习模块

- TextCNN [*Kim, 2014*]
- RCNN [*Xu+, 2016*]
- RNN+CNN+attention
  [*Vaswani+, 2017; Zhao+, 2019*]
- BERT+highway+GRU
  [*Devlin+, 2018; Srivastava+, 2015; Gulcehre+, 2014*]

改进的自适应参数DBSCAN
[*Dong+, 2018; Gao+, 2017; Zhu+, 2010*]

- 算法流程
- 正确性验证

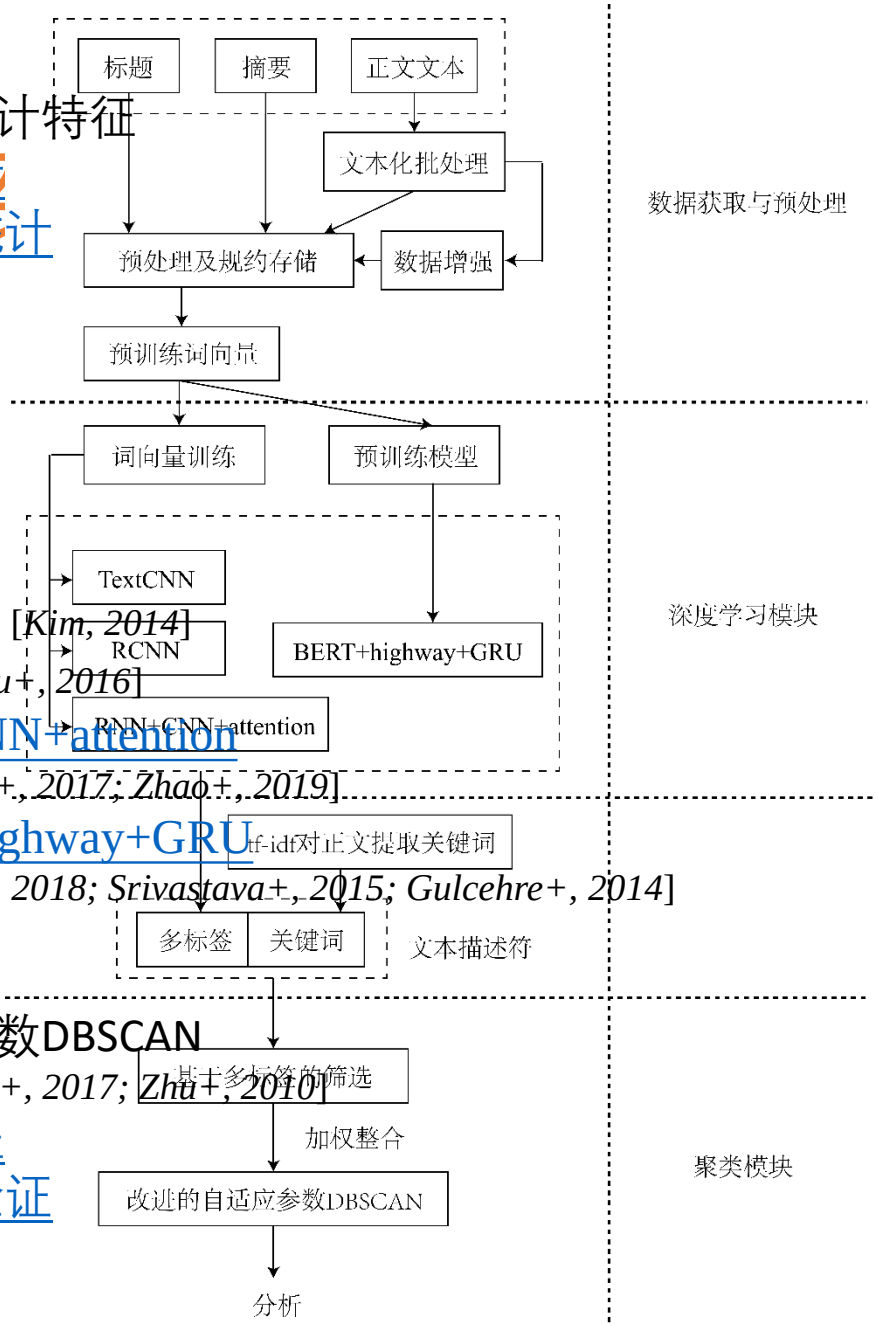


Figure 1. 算法整体框架

模型调优与结果分析

模型对比实验

- 预训练词向量
- 数据增强
- 四个模型的对比

参数调优

- Batch_size
- 学习率
- 训练方式

预测阈值设定

聚类结果分析

- 测试数据1
- 测试数据2

不足与展望

- **数据获取与预处理模块。**本课题使用统计方法确定各字段的截断词长度，按照本课题的方案，能含括84000条数据中的87.39%。这一步可以通过设置各字段长度为超参数，进一步完成优化，在保证稀疏度一定的程度上含括更多的数据。

- **深度学习模块。**由于算力有限，本课题使用的是BERT(FT+TM)的方式进行最终的训练。从实验结果可以看到，基于BERT预训练的方案比其他复合网络效果更好。因此，如果条件允许能使用我们爬取的科技文献数据进行预训练，或许能达到更好的分类效果。

此外，本课题在进行模型指标反馈时采用的top5，来计算预测多标签与原始label的精确度、召回率等。此处同样也可以设置为超参数，对合理范围内的topn进行讨论，使分类的反馈指标能更好的反应实际分类的效果。

最后，多标签预测向量的阈值选定。本课题只讨论了阈值为0.05、0.1、0.2和0.3四种情况，最终选择阈值为0.1。可以通过更小的步长，发现位于0.1左右某区间的最优阈值选择。

不足与展望

- **聚类模块。**由于本课题涉及的向量维度较大，算法需要计算点对的距离矩阵。为加快实验测试速度，也仅使用了200维向量。此外，在核密度估计时，需要计算百重积分优化核函数缩放参数，因此本课题通过计算机数值模拟在公式上做了近似处理。下一步可以研究其他的自适应聚类算法，以更精确的完成最优参数对的选择。

Thank you!

厚德博学 敬业乐群

分布统计

共计10个大类168个小类

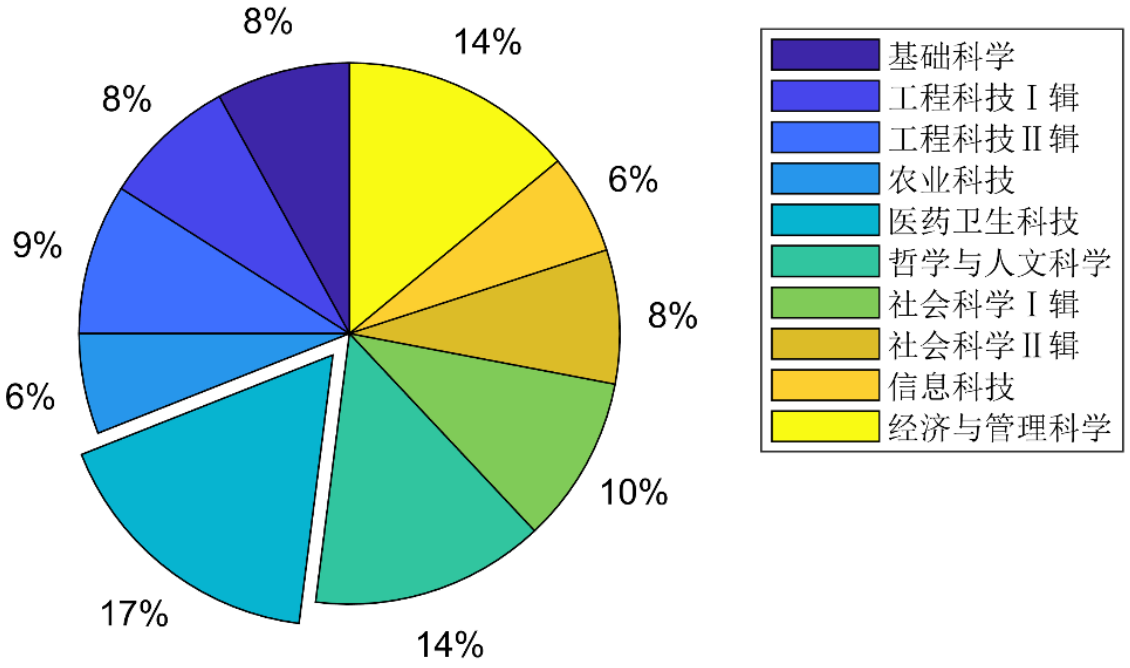


Figure 3. 爬取数据的分布统计图

词长度统计

表 1. 切词、分词后各字段词长度统计及 *Maxlen*

字段	最大长度	最小长度	众数	*Maxlen*
标题	47	1	7	12
摘要	813	2	75	128
正文	16501	2058	2664	4000

87.39%

数据增强

表 2(a)．基于过采样的数据增强方式总结

对象粒度	增强方式	操作	描述
Character	Random Aug	插入 替换 交换 删除	随机插入 替换 交换(打乱) 删除
	Keyboard Aug	替换	模拟键盘距离错误
Word	Random Word Aug	交换 删除	随机交换词 随机删除词
	WordEmbs Aug	插入 替换	从word2vec, GloVe或fasttext dictionary随机插入词 基于word2vec, GloVe或fasttext embeddings替换词
	Contextual WordEmbs Aug	插入 替换	基于BERT和XLNet语言模型插入词 基于BERT和XLNet语言模型替换词
Sentence	Contextual WordEmbs For Sentence Aug	插入	根据GPT2或XLNet prediction插入词

我爱北京邮电大学

$Maxlen = 5$

爱 北京 邮电 大学 空

表 2(b)．本课题所使用的数据增强示例

预处理后	数据增强方式		
	以“空”字符随机替换	打乱顺序	根据语言模型随机将部分词替换为最相似词
爱 北京 邮电 大学 空	空 北京 邮电 大学 空	邮电 大学 空 北京 爱	爱 北京 邮电 学院 空

Textcnn

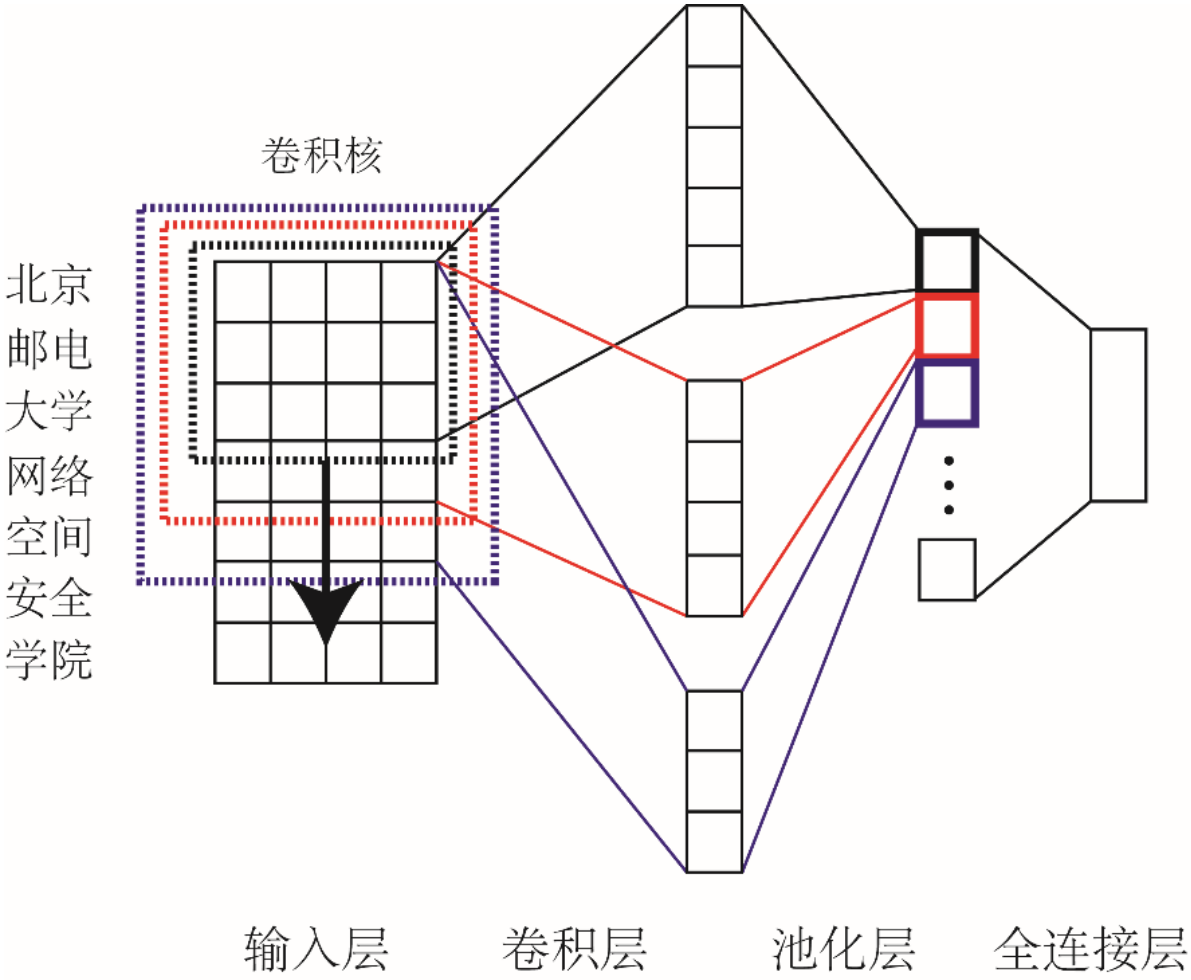


Figure 4. Textcnn结构简图

RCNN

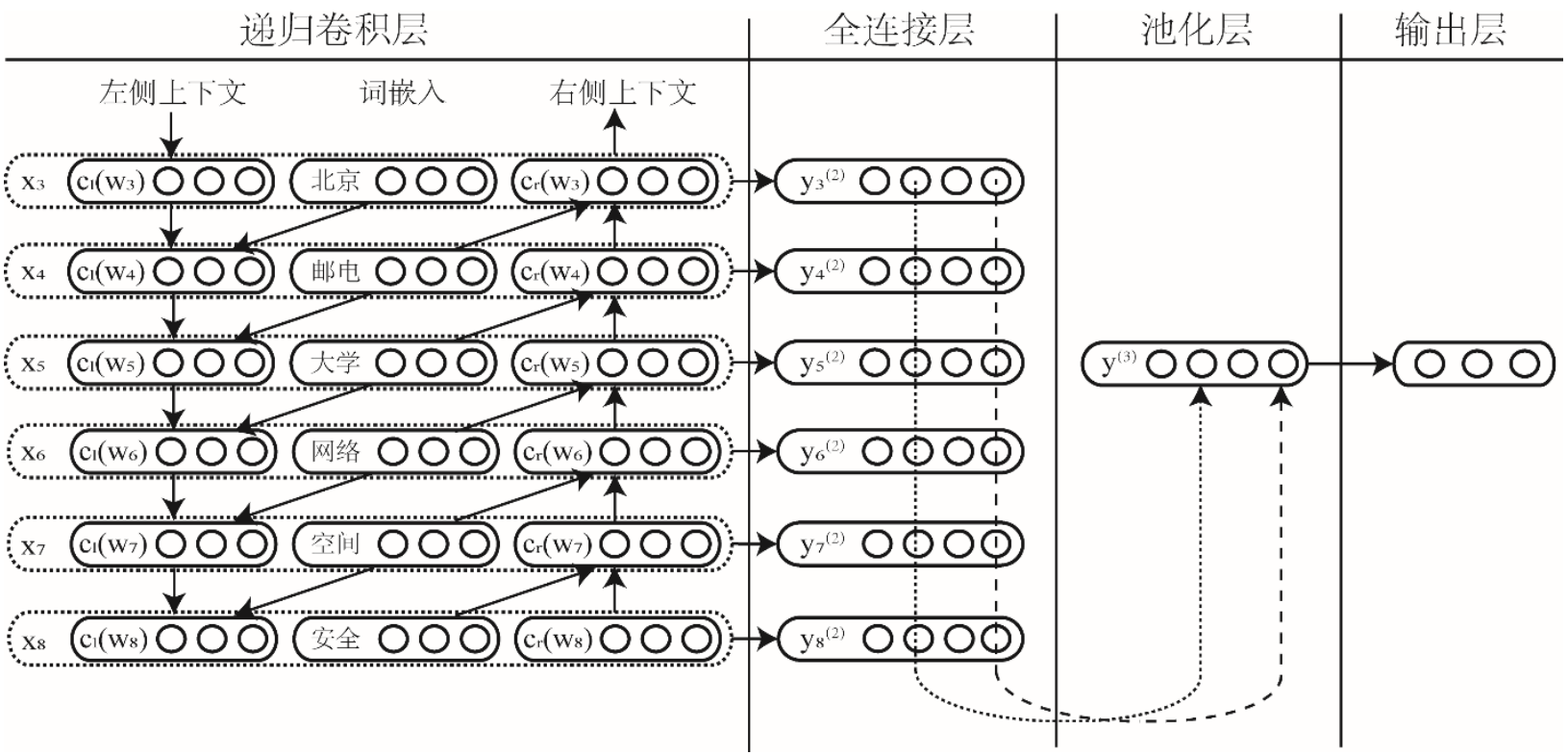


Figure 5. RCNN结构简图

CNN+RNN+attention

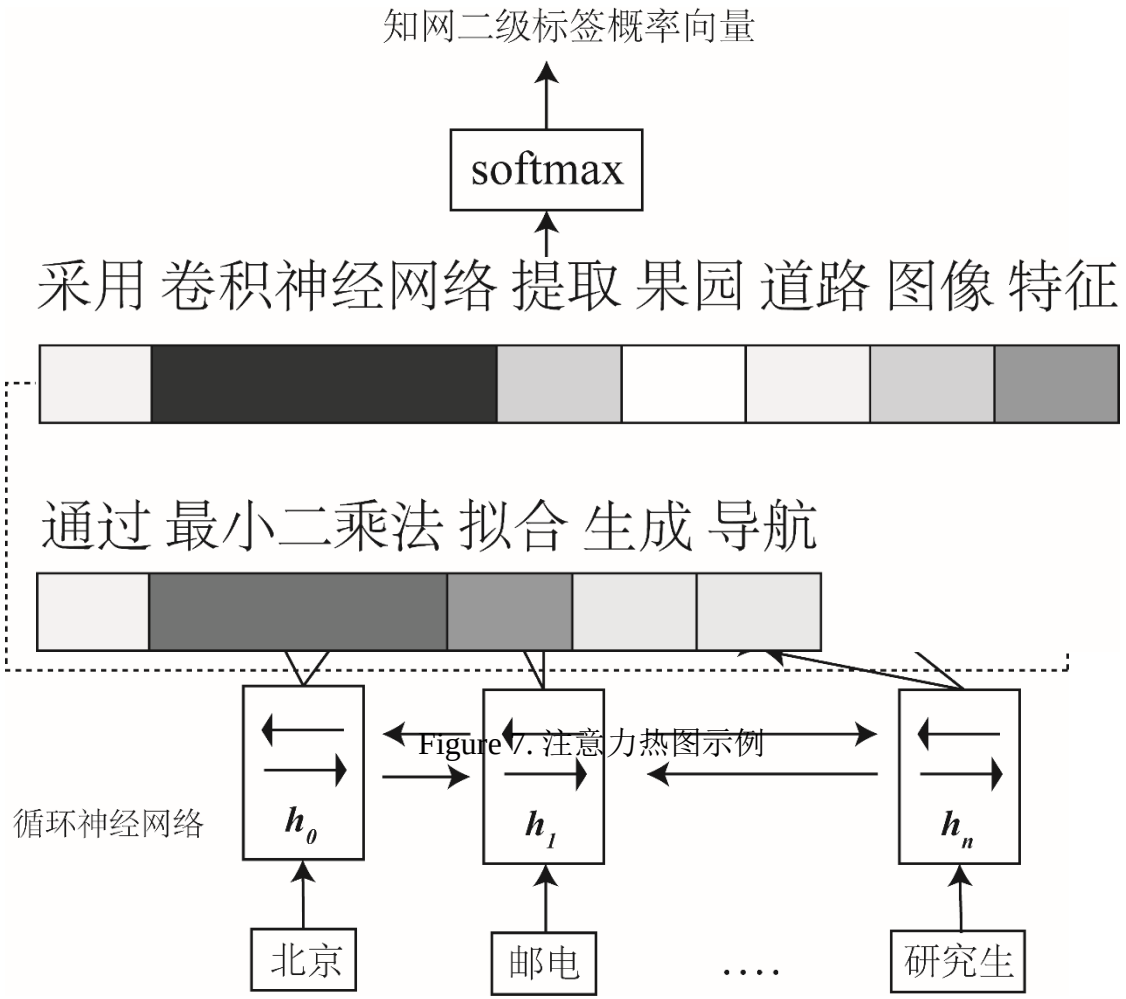
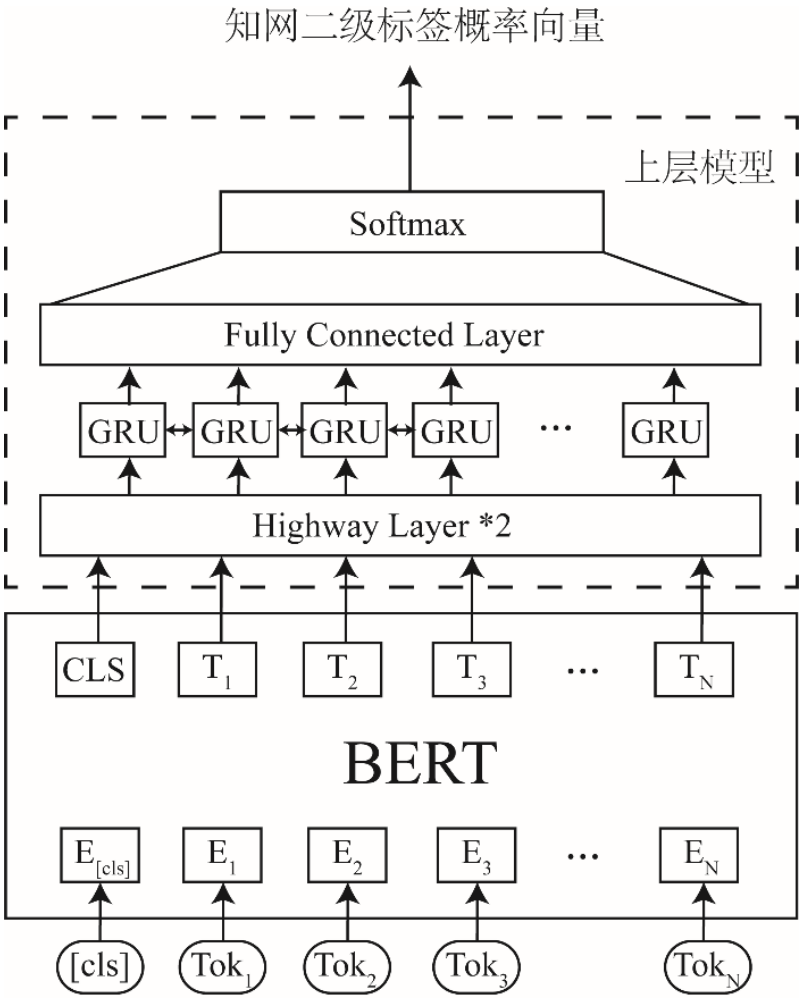


Figure 7. 注意力热图示例

Figure 6. 复合网络结构简图

BERT+highway+GRU



[*BERT-Base, Chinese*]

Figure 8. 基于BERT的分类网络简图

算法流程

表 3. 参数对的自适应算法步骤

输入：	数据集 D 包含 n 条整合后的文献描述符，维度为词向量维度 m 。
Step1:	对于 n 个点，计算点对之间的相互距离，得到距离矩阵 Dis 。
Step2：	根据公式 $\hat{f}_h^*(x)=n^{-\frac{4}{5}}\exp(-\frac{n^{\frac{1}{5}}}{2\delta^2})\sum_{i=1}^{n}K(x,x_i)$，结合距离矩阵 Dis，加和所有相同距离的 $\hat{f}_h^*(x)$ 值，得到距离-密度的图像。找到第一个峰值，确定 Eps 的合理取值范围；此后，遍历所有点，根据 Eps 通过期望计算，得到 $MinPts=\frac{1}{n}\sum P_i$ 。
Step3：	等步长对参数 Eps 区间进行划分，并依据不同参数对的值计算聚类中心。以所有聚类中心数的众数作为基准，筛除部分参数对。
Step4：	对 Step2 中剩余的每种情况，分别计算轮廓系数。
Step5：	选择 Step3 中最大的轮廓系数值所对应的参数对 $(Eps,MinPts)$ 。
输出：	$(Eps^*,MinPts)$

正确性验证

$Eps \in [1.8,2.2]$
$step = 0.05$
$E(Minpts) = 165$

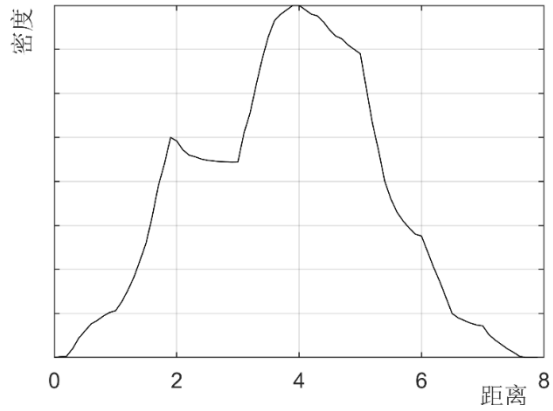


Figure9(b). 栾富庞伶讧砸宠Eps范围和$Minpts$期望

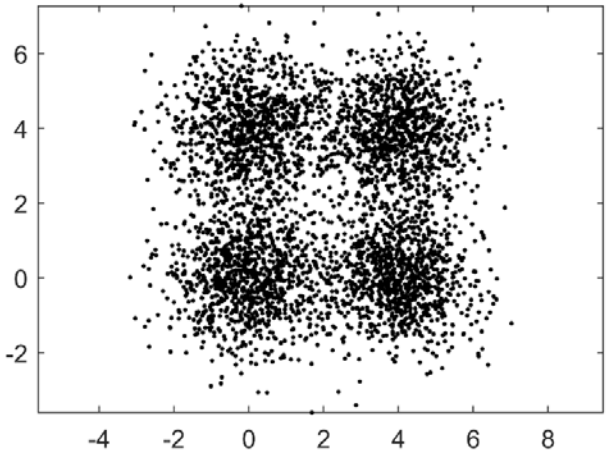


Figure 9(a). 产生4000个随机点

通过中心簇估算初步得到中心集合，再取轮廓系数最高的点作为中心点，对应的Eps作为最终的参数。

$C = \{(-0.001,0.010), (3.988,0.001), (0.001,4.000), (4.000,4.000)\}$
$Eps = 2.05$，$MinPts = 165$

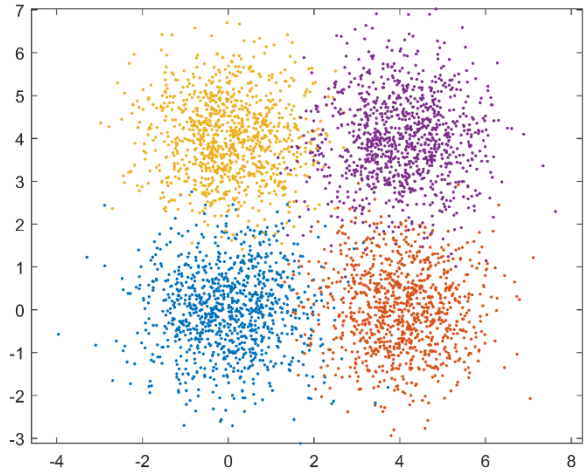


Figure9(c). 聊籽给枢

预训练词向量

- Wiki中文百科语料
-

表 4. 预训练词向量的效果测试

网络模型结构	评价指标 P / R / F_1	
	未使用预训练词向量	使用预训练词向量
TextCNN	0.8018 / 0.7966 / 0.7992	0.8056 / 0.7981 / 0.8018
RCNN	0.8113 / 0.8000 / 0.8056	0.8150 / 0.8029 / 0.8089
CNN+RNN+attention	**0.8195 / 0.8234 / 0.8214**	**0.8236 / 0.8277 / 0.8256**

数据增强

- 对占比不超过8%的数据进行增强
- “空”替换占原文比20%，基于语言模型的词换占原文比10%
-

表 5 数据增强的效果测试

网络模型结构	评价指标 P / R / F_1	
	未使用数据增强	使用数据增强
TextCNN	0.8056 / 0.7981 / 0.8018	0.8102 / 0.8024 / 0.8063
RCNN	0.8150 / 0.8029 / 0.8089	0.8213 / 0.8160 / 0.8186
CNN+RNN+attention	**0.8236 / 0.8277 / 0.8256**	**0.8281 / 0.8377 / 0.8329**

四个模型的对比

- 数据增强
- Batch_size=32
-

表 6 四种模型的效果对比

网络模型结构	评价指标 P / R / F_1	开始收敛轮数
TextCNN	0.8102 / 0.8024 / 0.8063	24
RCNN	0.8213 / 0.8160 / 0.8186	24
CNN+RNN+attention	0.8281 /0.8377 / 0.8329	25
BERT(FT+TM)	**0.8650 / 0.8555 / 0.8602**	**36**

参数调优

表 4. 预训练词向量的效果测试

网络模型结构	评价指标 P / R / F_1	
	未使用预训练词向量	使用预训练词向量
TextCNN	0.8018 / 0.7966 / 0.7992	0.8056 / 0.7981 / 0.8018
RCNN	0.8113 / 0.8000 / 0.8056	0.8150 / 0.8029 / 0.8089
CNN+RNN+attention	**0.8195 / 0.8234 / 0.8214**	**0.8236 / 0.8277 / 0.8256**

影响

收敛轮数
37
33
30
35

表 7(c) 训练方式和使用策略对模型的影响

训练方式	F_1 值	开始收敛轮数
NFT+TM	0.8359	24
FT+TM	**0.8740**	**33**
先联合训练后固定	0.8687	28

- Batchsize=256(single=64*4)
- 学习率=0.00005/动态优化

预测阈值设定

-
-
-

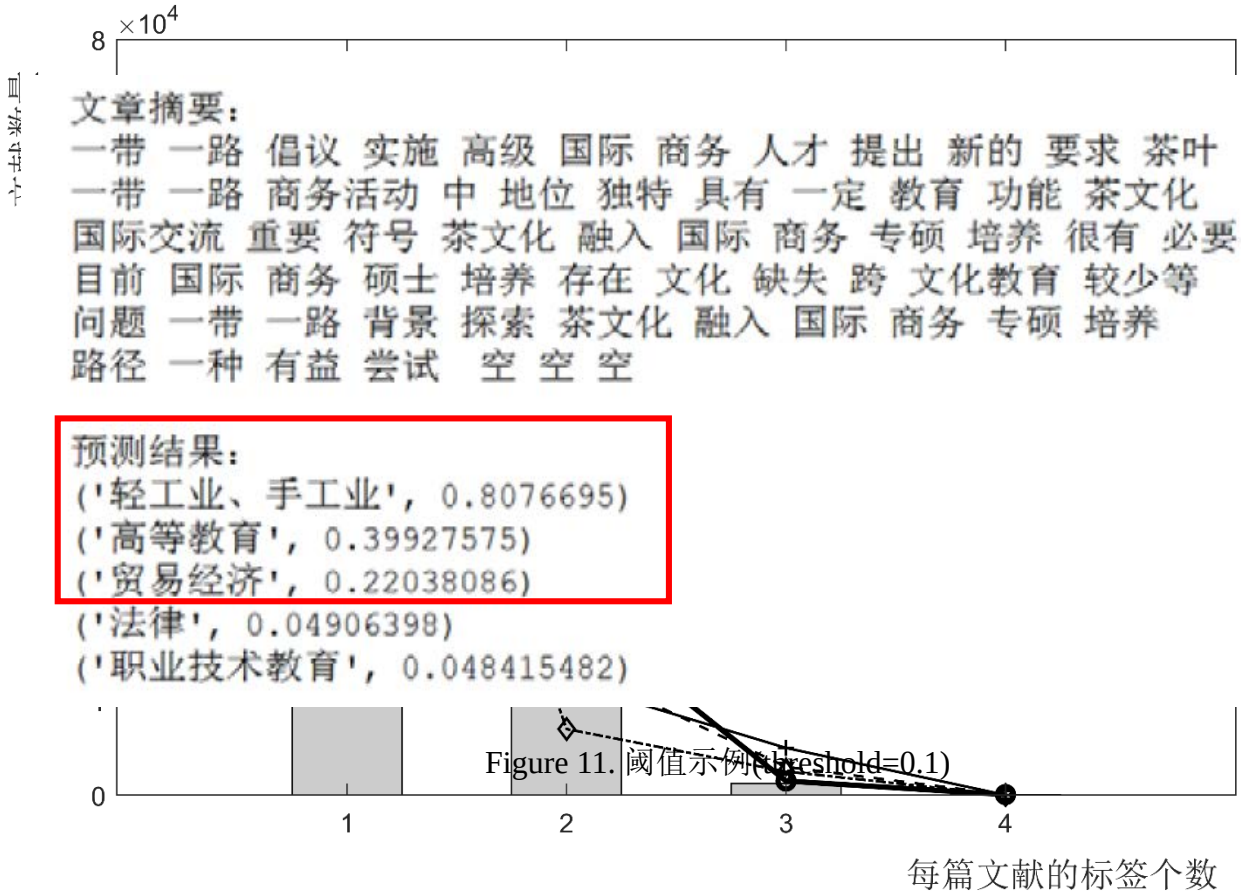


Figure 11. 阈值示例(threshold=0.1)

Figure 10. 不同阈值下各标签组合的数量统计

聚类结果分析-测试数据1

- 随机抽取2017年9月-12月知网文献2000篇，经过多标签删选和相似度判定，共78篇进行聚类。
- 聚类主要分为三个簇，最大簇含25篇，以其中两篇为例：

表 8(a) 第一批数据分析(文献数据)

文献摘要	预测多标签
“4G 网络时代手机自媒体迅速发展,传统茶文化不断发扬光大。本文探究手机自媒体与传统茶文化在思政教育上创新结合的价值与路径,推动思想政治教育模式不断创新。”[1]	轻工业 手工业 高等教育 信息与知识传播
“茶艺是一门应用技能型学科,是茶文化的重要组成部分。在"一带一路"背景下,高校茶艺课程开展双语教学,达到了传播中国茶文化,发展中国茶艺技能的目的。基于学科特点和对高素质人才的需求,对茶艺课程的教学内容、教学方法、考核方式等方面进行了教学改革与探索。”[2]	轻工业 手工业 高等教育

[1]褚洪彦. 4G网络时代手机自媒体与传统茶文化在思想政治教育上的创新结合[J]. 福建茶叶, 2017, 0(12): 240–240.
[2]王丽, 叶国盛. “一带一路”背景下高校茶艺课程双语教学模式初探[J]. 海峡教育研究, 2017(03): 61–65.

聚类结果分析-测试数据2

- 限定发表时间为2020年2月5日，且全文检索内含“冠状病毒”字样，共爬取数据218条。经过上述处理后与筛选后，共有46篇进入聚类算法。其结果主要分为五个簇，最大簇包含文本10篇。以其中两篇为例：

表 8(b) 第二批数据分析(报纸数据)

正文快照(摘要)	预测多标签
“青海一新型冠状病毒感染的肺炎患者因隐瞒行程被警方立案，系全国首例。目前，多地发布通告强调，明知已感染或可能感染新型冠状病毒，故意进入公共场所或隐瞒情况与他人接触，构成犯罪的一律追究刑事责任。”[3]	病毒学 法律学
“随着“新型冠状病毒肺炎”的发现、传播，患者或疑似症状者纷纷被隔离，那么，其中的劳动者与用人单位的劳动关系如何处理？被隔离后用人单位能否停发工资传染病防治法第三条规定。”[4]	病毒学 法律学

[3] 杨玉龙. 因隐瞒行程被警方立案具有警示意义[J]. 中华工商时报, 2020: 003.
[4] 颜梅生. 因疫情被隔离，事关劳动关系的那些事儿[J]. 中国妇女报, 2020: 006.

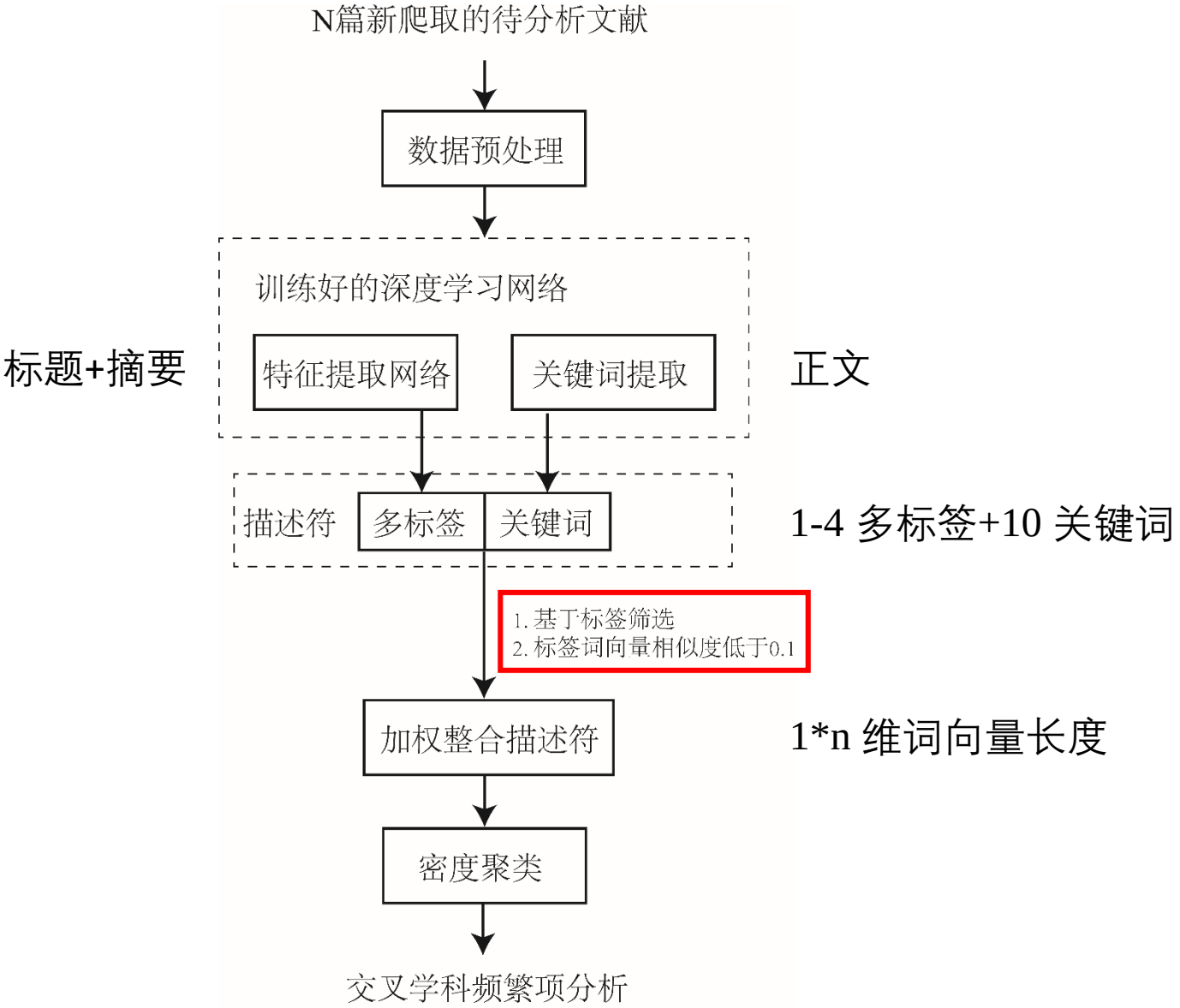


Figure 2. 不同阈值下各标签组合的数量统计