

基于CCSS十年追踪调查的学生留言文本分析

刘哲¹, 单剑锋¹, 杨立军²

(1. 南京邮电大学 电子与光学工程学院、柔性电子(未来技术)学院;
2. 南京邮电大学 教学质量监控与评估中心, 江苏 南京 210023)

摘要: 通过调查问卷定期收集大学生的意见是高校提高办学质量的重要举措, 问卷中的留言在很大程度上可以体现出学生的诉求、期待与评价。然而, 学生留言往往是只言片语, 存在词汇稀疏、语义特征匮乏等问题。为此, 提出一种改进的Mini-Batch K-Means聚类算法, 通过使用聚类方法将短文本按照相似度大小变换成长文本作为LDA模型的输入。改进算法每次迭代只使用部分数据, 通过随机抽样来处理大规模数据集, 在抽样样本批次内找出新的聚类中心, 然后将该聚类中心更新到全局聚类中心中, 有助于稳定聚类中心的更新, 并提供更准确的聚类结果。实验结果表明, 改进聚类方法所需运行时间可维持在10 s以内, 模型平均PMI值最高提升了15.1%。文本分析结果显示, 10年来, 大学生在就业机会、专业、社会实践、课程、教育资源等方面表现出较为稳定的关注程度, 个别年份突出关注宿舍等个性化问题。

关键词: 学生留言; Mini-Batch K-Means 聚类算法; 主题模型; LDA; TF-IDF

DOI: 10.11907/rjdk.241486

开放科学(资源服务)标识码(OSID):

中图分类号: TP391; G647

文献标识码: A

文章编号: 1672-7800(2025)009-0034-07



Text Analysis of Student Message Based on CCSS Ten-year Follow-up Survey

LIU Zhe¹, SHAN Jianfeng¹, YANG Lijun²

(1. School of Electronic and Optical Engineering and School of Flexible Electronics (Future Technology), Nanjing University of Posts and Telecommunications; 2. Teaching Quality Monitoring and Evaluation Center, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Regularly collecting opinions from college students through survey questionnaires is an important measure for universities to improve the quality of education. The comments in the questionnaire can largely reflect students' demands, expectations, and evaluations. However, student messages are often fragmented, with issues such as sparse vocabulary and lack of semantic features. To this end, an improved Mini-Batch K-Means clustering algorithm is proposed, which transforms short texts into long texts according to their similarity using clustering methods as inputs for the LDA model. The improved algorithm only uses partial data in each iteration, processes large-scale datasets through random sampling, finds new cluster centers within the sampling batch, and then updates the cluster centers to the global cluster centers, which helps to stabilize the update of cluster centers and provide more accurate clustering results. The experimental results show that the improved clustering method can maintain a running time of less than 10 seconds, and the average PMI value of the model has increased by up to 15.1%. The text analysis results show that in the past 10 years, college students have shown a relatively stable level of attention in terms of employment opportunities, majors, social practice, courses, educational resources, etc., with individual years highlighting personalized issues such as dormitories.

Key Words: student message; Mini-Batch K-Means clustering algorithm; thematic model; LDA; TF-IDF

收稿日期: 2024-05-28

扫描二维码阅读全文:

基金项目: 国家社会科学基金一般项目(22BTJ030)

作者简介: 刘哲(1999-), 男, 南京邮电大学电子与光学工程学院、柔性电子(未来技术)学院硕士研究生, 研究方向为计算机视觉和深度学习; 单剑锋(1967-), 男, 博士, 南京邮电大学电子与光学工程学院、柔性电子(未来技术)学院副教授, 研究方向为智能信号处理; 杨立军(1969-), 女, 硕士, 南京邮电大学教学质量监控与评估中心研究员, 研究方向为教育质量评价、大学生学习与发展。



0 引言

通过调查问卷定期收集大学生的意见是高校提高办学质量的重要举措,也是高校以学生为中心办学理念的体现。问卷调查中的留言在很大程度上可以反映出学生的诉求、期待和评价,对留言的内容进行分析已经成为建设持续改进的高校质量保障机制的基础工作。目前国际上大学生问卷调查项目主要包括大学生就读经验调查(College Student Experience Questionnaire, CSEQ)、美国大学生学习投入调查(National Survey of Student Engagement, NSSE)等。国内的中国大学生学习与发展追踪研究(China College Student Survey, CCSS)是清华大学教育研究院在引介NSSE的基础上于2009年发起的全国性大学生学情调查项目。该项目通过调查问卷的形式来了解大学生的学习状态与体验,问卷的最后一部分用于收集学生对学校的意见与建议等信息。

随着国内CCSS项目的深入开展,很多高校已经采集了10余年的问卷调查数据,积累了大量文本信息,特别是学生评论与留言部分,不仅是学生心声的直接反映,更是洞察教育现状、优化管理策略的宝贵资源。采用合适的方法对这些文本信息进行高效识别与精准分析能够深入挖掘学生反馈中体现出的问题、不满与期望,如课程设置合理性、教学质量满意度、学生支持服务需求等,有助于学校及时发现教学管理中的薄弱环节与潜在问题,建立精细化、个性化、科学化的学生管理体系,实现教育质量的动态提升与持续改进^[1-4]。

1 相关研究

目前,LDA(Latent Dirichlet Allocation)主题模型已广泛应用于文本分析领域,其是包括词项、主题和文档3层结构的文档生成概率模型,能够发现潜在主题^[5]。然而,未经修改的LDA主题模型一般适用于新闻等经过初次加工的语料,而学生留言是短文本,语义信息稀疏、噪声大、文本语言口语化,极大增加了对其进行主题挖掘的难度^[6-7]。为解决这一问题,国内外学者对LDA模型进行了大量改进。例如,Zhao等^[8]提出Twitter-LDA模型,其认为每个用户存在1个用户—主题矩阵,从该矩阵中生成主题,从而展现出不同用户的关注点。黄波等^[9]提出基于向量空间模型(Vector Space Model, VSM)与LDA模型的微博话题发现算法,该算法将LDA模型提取的文档间语义信息与基于VSM的词频—逆文件频率(Term Frequency-Inverse Document Frequency, TF-IDF)权重词向量融合在一起,用于计算文本相似度。Apaza等^[10]利用LDA模型从大学课程大纲、在线课程大纲或课程介绍中提取信息,并结合评分信息为用户进行课程推荐。Moody^[11]在Word2vec中

skip-gram方法的基础上构建了一个LDA2Vec模型,同时考虑文档向量和单词向量以提高生成词向量的质量。

传统的K-Means算法因步骤简单、能有效处理大数据集而被广泛使用,应用于短文本聚类时更是简单而有效^[12]。然而,该算法一旦遇到聚类簇数较大的情况时运行时间便会大大增加。众多学者针对这一不足提出了大量优化与改进策略。例如,左进等^[13]通过优化选择 K 个初始化中心的过程,使得聚类结果更加接近真实聚类中心,减少了迭代次数,提高了聚类质量和异常检测率。孙志鹏等^[14]提出一种基于加权距离计算的自适应粗糙K-Means算法,有效提高了聚类准确度和执行效率,适用于高维数据的处理。Huang等^[15]提出一种特征值加权的K-Means算法,通过不同权重在迭代过程中选择中心点 K 值。周本金等^[16]提出一种最大化减少当前误差平方和的K-Means初始聚类中心优化方法,能有效减少聚类迭代次数,提高聚类质量。

学生问卷中包含的数据十分庞大且具有随机性,LDA模型挖掘学生留言短文本中蕴含主题的难度较大,需先对短文本进行聚类。因此,本文提出将Mini-Batch K-Means聚类算法加入到融合TF-IDF的LDA模型中,算法每次迭代只使用部分数据,通过随机抽样或分块方式来处理大规模数据集,在部分样本内进行找出新的聚类中心后,将当前样本批次的聚类中心更新到全局聚类中心中,以便在较大的数据集中提高模型的点互信息(Pointwise Mutual Information, PMI)值。

2 聚类算法与主题模型

2.1 聚类算法

聚类算法根据用户要求的聚类数给出一个初步聚类方法,通过不断迭代去遍历数据集中的样本对象,改变上一次对类的划分结果,达到每次迭代改进后的聚类结果都比上一次更接近理想结果的目的。理想结果即为聚集在一起的数据相似程度越高,离得较远的数据相似度越低。

2.1.1 K-Means 聚类算法

K-Means是一种常用的无监督学习算法,用于将数据集中的样本划分为 K 个不同的簇。其是一种迭代的聚类算法,通过最小化数据点与所属簇中心之间的距离来确定簇的边界,在满足人为设置好的终止条件后结束迭代。迭代终止条件可参考以下3点设置^[17]:①所有对象被重新分配到不同的类;②聚类中心停止改变或在一个极小的范围内微动;③满足优化函数的规定或迭代次数已封顶。

K-Means算法的步骤为:

- (1)选择聚类簇的数量 K 。
- (2)随机选择 K 个数据点作为初始的聚类中心。
- (3)对于每个数据点,计算其与各个聚类中心之间的距离,通常使用欧氏距离或曼哈顿距离。欧氏距离的计算

公式为:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

式中: x 和 y 为两个数据点; n 为数据点的维度。

曼哈顿距离的计算公式为:

$$d(x, y) = \sum_{i=1}^n |x_i - y_i| \quad (2)$$

式中: x 和 y 为两个数据点; n 为数据点的维度。

(4) 将每个数据点分配给距离最近的聚类中心, 形成 K 个簇。

(5) 对于每个簇, 计算簇中所有数据点的均值, 将均值作为新的聚类中心。

(6) 重复步骤 4 和步骤 5, 直到聚类中心不再发生变化或达到最大迭代次数。

2.1.2 Mini-Batch K-Means 聚类算法

Mini-Batch K-Means 算法是对传统 K-Means 算法的一种改进, 主要体现以下 3 个方面: ①在数据处理方式上, 传统算法使用整个数据集来计算聚类中心和更新簇分配, 每次迭代都对所有数据点进行处理; 改进算法的每次迭代只使用部分数据, 通过随机抽样或分块方式来处理大规模数据集; ②在计算效率方面, 传统算法由于每次迭代都需要处理整个数据集, 面对大规模数据集时计算复杂度较高, 消耗的时间和内存资源也较多; 改进算法通过使用部分数据来更新聚类中心, 减少了计算量, 提高了算法效率; ③在收敛速度方面, 传统算法由于每次迭代都使用整个数据集, 可能需要较多的迭代次数才能收敛到最终结果; 改进算法使用部分数据进行更新, 算法可以更快收敛。然而, 由于仅使用部分样本批次, Mini-Batch K-Means 算法的结果可能略有偏差。

Mini-Batch K-Means 聚类算法流程如图 1 所示。具体步骤为:

(1) 读取数据集, 输入聚类个数 K 值。

(2) 选择 K 个初始聚类中心点, 设置批处理大小 (mini-batch size) 参数, 用于决定每次迭代中用于更新聚类中心的样本数量。通过每次从整个数据集中随机选择一部分样本用于批处理, 可以引入随机性, 使得算法更具鲁棒性。此外, 批处理的随机选择还有助于保持对数据集的代表性, 避免在每次迭代中只使用部分数据而导致的偏差。

(3) 随机选择一个批处理大小的样本批次, 并在当前样本批次内计算每个数据点与各个聚类中心之间的距离, 将数据点分配给距离最近的聚类中心。

(4) 在当前样本批次内计算簇中所有数据点的均值, 并将均值作为新的聚类中心。

(5) 在当前样本批次内找出新的聚类中心后, 将该聚类中心更新到全局聚类中心中, 一般使用指数移动平均 (Exponential Moving Average) 方法, 有助于聚类中心的稳定更新, 并提供更准确的聚类结果。

(6) 重复执行步骤 3、4、5, 直至达到预定的迭代次数或收敛条件, 然后输出聚类结果。

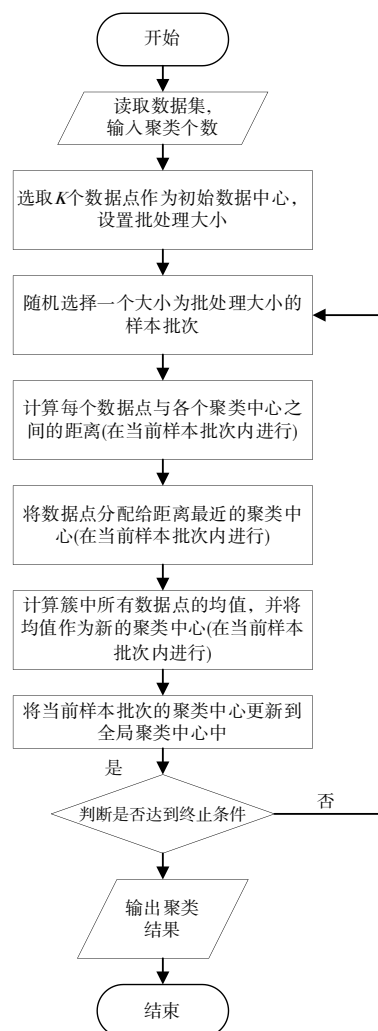


Fig. 1 Mini-Batch K-Means algorithm flow

图 1 Mini-Batch K-Means 算法流程

2.2 主题模型

2.2.1 LDA 主题模型

LDA 主题模型在文档—单词关系中引入主题的概念, 形成了文档—主题—词的 3 层贝叶斯模型^[18]。在该模型中, 文本 m 可以表示成多个主题的联合分布, 记为 $P(z)$, 每个主题又是词汇表中所有单词的概率分布, 记为 $P(w|z)$ 。因此, 文本中每个单词的概率分布为:

$$p(w_i) = \sum_{j=1}^K p(w_i|z_i = j) p(z_i = j) \quad (3)$$

式中: K 为主题个数; $i \in [1, N_m]$, N_m 为数据集中所有单词的个数。

LDA 模型结构如图 2 所示。其中, M 表示语料中文档的个数; K 表示语料中主题的个数, 实际训练中 K 值可作为超参数进行调整; θ_m 为第 m 篇文档的主题分布, 为多项式分布; α 为文档—主题分布的 Dirichlet 先验分布; φ_k 为第 k 个主题的词汇分布, 亦为多项式分布; β 为主题—词汇分布的 Dirichlet 先验分布; $Z_{m,n}$ 表示第 m 篇文档中第 n 个词汇所属的主题; $W_{m,n}$ 表示第 m 篇文档的第 n 个词汇。模型中文档与文档之间、主题与主题之间相互独立。

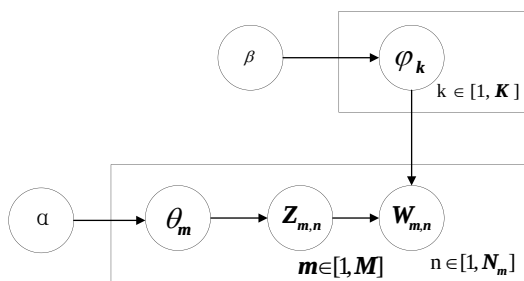


Fig. 2 LDA model structure

图2 LDA模型结构

由图2可知, LDA模型的具体生成过程为: ①生成主题—词汇分布, 根据第 k 个主题的Dirichlet先验分布生成第 k 个主题的词汇分布 φ_k ; ②生成文档—主题分布, 根据第 m 篇文档的Dirichlet先验分布生成第 m 篇文档的主题分布 θ_m ; ③根据主题分布 θ_m 得到1个主题编号 k , 再根据主题 k 的词汇分布 φ_k 生成词 $w_{m,n}$ 。

2.2.2 吉布斯抽样法

吉布斯抽样法是一种简单的马尔可夫链蒙特卡洛算法, 经常用于进行LDA模型概率推导。该方法的思想为对于已知概率分布 $\pi(x)$, $x = (x_1, x_2, \dots, x_n)$, 如果 π 太过复杂以致不能直接从其中抽样, 可以通过构造一个非周期且不可约的马尔可夫链来间接获取样本。由于模拟的数值可以被视作来自目标分布的独立样本, 当马尔可夫链足够长时, 便可以用其稳态分布来推断 π 的重要特征。

2.2.3 TF-IDF算法

TF-IDF算法是一种常用的文本特征提取方法, 用于衡量一个词对于一个文档集合中某个文档的重要性。其中, TF指一个词在文档中出现的频率, 表示一个词在文档中的重要程度, 通常使用词频除以文档中的总词数来计算; IDF指一个词在整个文档集合中的稀有程度, 用于降低常见词对文档的重要性, 增加罕见词的重要性, 通常使用总文档数除以包含该词的文档数的对数来计算。TF-IDF算法涉及的公式包括:

$$TF(w, d) = (\text{词}w\text{在文档}d\text{中出现的次数}) / (\text{文档}d\text{中的总词数}) \quad (4)$$

$$IDF(w) = \log((\text{总文档数}) / (\text{包含词}w\text{的文档数})) \quad (5)$$

$$TF-IDF(w, d) = TF(w, d) * IDF(w) \quad (6)$$

TF-IDF算法的具体步骤为: ①对于给定的文档 d , 计算文档中每个词 w 的TF值; ②对于每个词 w , 计算其IDF值; ③将TF值与IDF值相乘, 得到每个词 w 在文档 d 中的TF-IDF值。

2.2.4 评价指标

主题连贯性用于衡量主题中词语之间的关联性和连贯性, 即主题中的词语是否具有一定的语义相关性和组织结构, 是评估LDA模型生成主题质量的重要指标之一。连贯的主题可以提供更好的解释性, 有助于理解主题模型学习到的语义信息。常见的主题连贯性评估指标之一为PMI。PMI值越高, 表明模型性能越好。计算公式为:

$$PMI(x, y) = \log(P(x, y) / (P(x) * P(y))) \quad (7)$$

式中: $P(x, y)$ 为词语 x 和词语 y 共同出现的概率; $P(x)$ 和 $P(y)$ 分别为词语 x 和词语 y 独立出现的概率。

3 实验方法与结果分析

3.1 实验环境与参数设置

实验硬件系统为NVIDIA GeForce RTX 3060 Laptop GPU, 软件系统环境为Windows 11, 开发语言为Python, 版本为3.9.11, 开发软件为PyCharm。

在训练K-Means聚类算法模型时, 取参数 $n_{init} = 10$, 表示运行K-Means算法的次数, 每次运行使用不同的初始聚类中心。 t 收敛判断的阈值为 $1e-4$, 每次运行K-Means算法的最大迭代次数为300。在训练Mini-Batch K-Means聚类算法时, 每个批次的样本数量 $batch_size = 100$, 参数 $n_{init} = 3$, 每次运行K-Means算法的最大迭代次数为300。在训练LDA模型时, 设置主题数范围为2~20, α 和 η 参数值为LDA模型根据语料库自动设置。聚类簇数 k 值分别设置为10、100、200、300、400、1 000。

3.2 数据来源与预处理方法

本文采用2011~2021年CCSS项目调查问卷中包含的大学生留言数据, 共计33 403条学生评论, 将其保存在Python文件中作为预处理数据。首先使用结巴分词器对数据进行处理, 其次使用停用词表去除无意义的字以及各种符号^[19]。所用停用词来自哈尔滨工业大学开源的停用词表^[20]。

3.3 实验结果与分析

3.3.1 比较实验

比较k-medoids聚类算法、K-Means聚类算法与本文提出的Mini-Batch K-Means聚类算法的运行时间随簇值 k 的变化情况^[21]。结果见表1。可以看出, 不论簇值 k 如何变化, Mini-Batch K-Means聚类算法所需的运行时间总是少于另外两种算法。

Table 1 The variation situation of running time of three clustering algorithms with cluster value k

表1 3种聚类算法运行时间随簇值 k 的变化情况 (s)

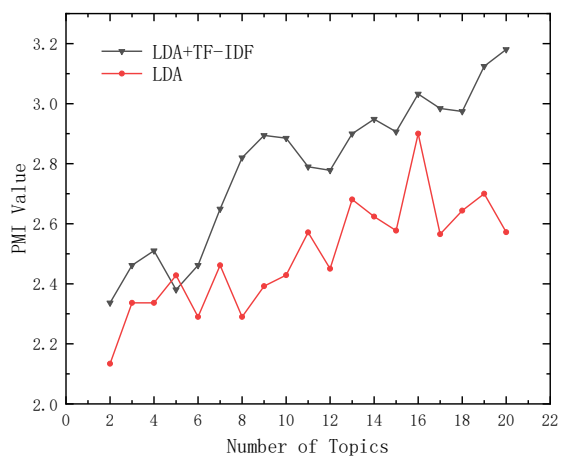
簇值 k	运行时间		
	K-Means 聚类 算法	k-medoids 聚类 算法	Mini-Batch K-Means 聚类算法
10	2.52	30.17	0.85
100	16.08	23.95	2.38
200	30.89	21.36	3.31
300	56.00	24.36	5.47
400	95.41	22.10	7.92
1 000	193.51	22.04	9.58

本文在验证时选取的簇类值共有5种, 为考察模型的适用性, 选取最小和最大两种 k 值的结果进行展示。由于1个主题个数对应1个PMI值, 本文实验主题数量为2~20, 将PMI总值除以19变为平均PMI值来进行比较。当 $k=10$

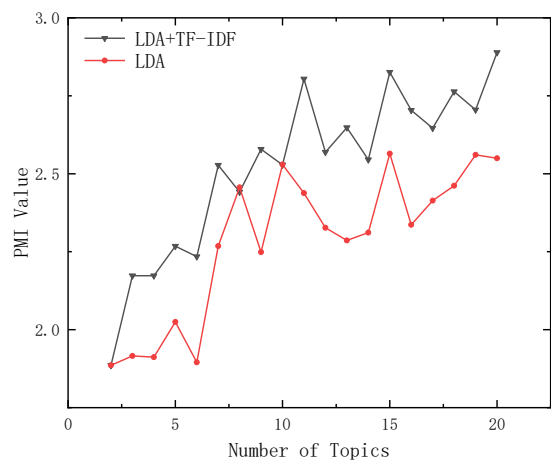
和 $k=1\ 000$ 时,3种算法分别融合在LDA模型和LDA+TF-IDF模型时PMI值随主题个数变化的情况如图3和图4所示。可以看出,当使用Mini-Batch K-Means聚类算法时,两种 k 值下LDA+TF-IDF模型中的平均PMI均大于使用K-Means和k-medoids算法时。此外,在使用相同聚类算

法时,LDA+TF-IDF模型的平均PMI值均大于LDA模型。

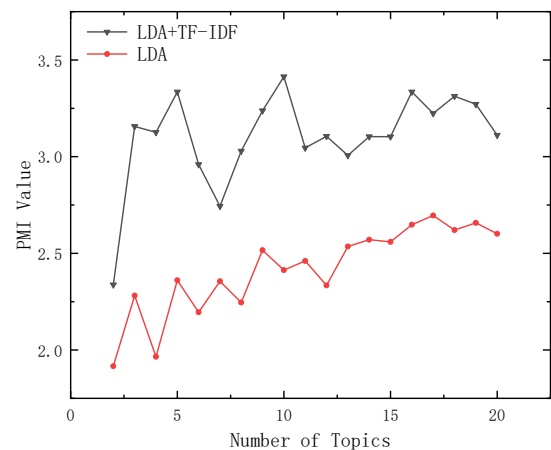
如表2所示,3种算法融合LDA+TF-IDF模型后的平均PMI值会随着簇值 k 的变化而变化,且使用Mini-Batch K-Means聚类算法的模型平均PMI值总是大于其他两个聚类算法,展现出更好的聚类效果。



(a) K-Means



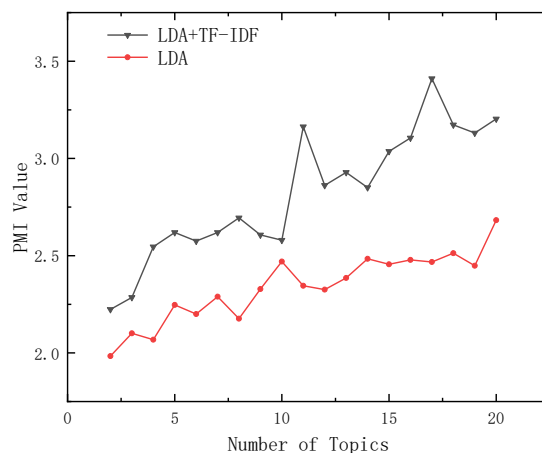
(b) k-medoids



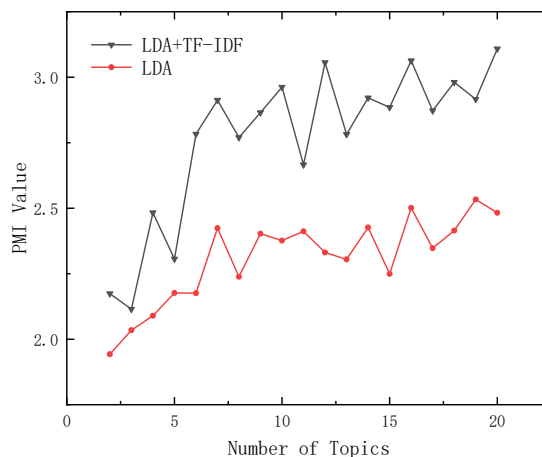
(c) Mini-Batch K-Means

Fig. 3 The variation of PMI values of three algorithms with the number of topics when $k=10$

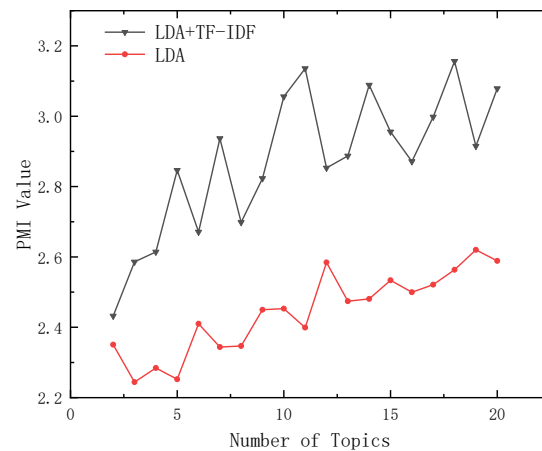
图3 $k=10$ 时3种算法的PMI值随主题个数变化的情况



(a) K-Means



(b) k-medoids



(c) Mini-Batch K-Means

Fig. 4 The variation of PMI values of three algorithms with the number of topics when $k=1\ 000$

图4 $k=1\ 000$ 时3种算法的PMI值随主题个数变化的情况

Table 2 The variation of average PMI values with cluster value k after the fusion of LDA+TF-IDF models using three algorithms

表 2 3 种算法融合 LDA+ TF-IDF 模型后的平均 PMI 值随簇值 k 的变化情况

簇值 k	平均 PMI 值		
	K-Means 聚类算法	k-medoids 聚类算法	Mini-Batch K-Means 聚类算法
10	2.78	2.53	2.96
100	2.80	2.75	2.93
200	2.91	2.95	3.06
300	3.09	2.77	3.19
400	2.84	2.81	2.88
1 000	2.85	2.84	2.88

3.3.2 验证实验

选取 PMI 值最高的模型进行验证实验,即将 Mini-Batch K-Means 聚类算法中的 k 值设置为 300。对经过 Mini-Batch K-Means 聚类后生成的新文本进行分析,在 Python 中导入 matplotlib.pyplot 和 jieba 库,将文本中的特殊字符转换为空格。创建一个空字典 counts,遍历分词列表 words,统计每个词语出现的次数,将词语及其出现的次数存储在字典 counts 中。按照出现次数对字典 counts 中的词语进行降序排列,排序结果存储在列表 item-s 中,输出出现次数最多的词语及其频数。本文研究对象为学生对学校的留言,因此在输出关键词时去掉了一些出现次数多但对研究没有用处的词语,如学校、学生等。2011-2021 年间出现频次最高的 7 个关键词如表 3 所示。

Table 3 7 keywords with the highest frequency from 2011 to 2021
表 3 2011-2021 年间出现频次最高的 7 个关键词

年份	关键词(按照频次由高到低排序)
2011 年	就业机会、教育资源、国家政策、专业、社会实践、课程、奖学金
2012 年	就业机会、国家政策、社会实践、专业、教育资源、老师授课、课程
2013 年	就业机会、社会实践、教育资源、专业、国家政策、课程、老师授课
2014 年	就业机会、专业、社会实践、教育资源、课程、国家政策、老师授课
2015 年	就业机会、社会实践、专业、国家政策、课程、教育资源、奖学金
2016 年	就业机会、专业、教育资源、课程、国家政策、社会实践、奖学金
2017 年	就业机会、专业、课程、社会实践、教育资源、国家政策、奖学金
2018 年	就业机会、专业、教育资源、奖学金、课程、社会实践、国家政策
2019 年	食堂、就业机会、专业、课程、社会实践、国家政策、奖学金
2020 年	就业机会、专业、课程、奖学金、社会实践、国家政策、宿舍
2021 年	就业机会、专业、课程、国家政策、社会实践、教育资源、奖学金

可以看出,就业机会一直是大学生们最为关注的话题,还有一些方面是其关注的重点,如社会实践、教育资源、国家政策、专业、课程、食堂、奖学金、老师授课和宿舍等。具体来说,大学生在教育资源方面的诉求为:①学校资源可以更加均衡些,减少差异;②学校可以在科研经费、教学设备上增加投入;③资源投入与分配可以更加公平、公正、透明。在国家政策方面的诉求为:①加大对贫困学生的支持;②出台相关政策使学生更加全面发展,锻炼实践能力;③落实有助于大学生就业的国家政策,提高就业福利。在课程方面的诉求为:①使学生可以更加自主地选择感兴趣的课程;②减少一些与专业无关的课程

安排;③将课程学习更多地与实际相结合,与就业要求相联系。在专业方面的诉求为:①适当降低学生专业调剂的难度,使其可以选择自己感兴趣的专业;②提高学生的专业技能,增强实践能力。在老师授课方面的诉求为:①老师在授课时多调动学生的积极性,多与学生进行交流;②老师在授课方面多花费一些精力。

图 5 为 2011-2021 年间所有留言汇总起来的关键词统计结果。可以看出,前 10 位从高到低依次为就业机会、专业、社会实践、教育资源、课程、国家政策、食堂、奖学金、老师授课和宿舍。

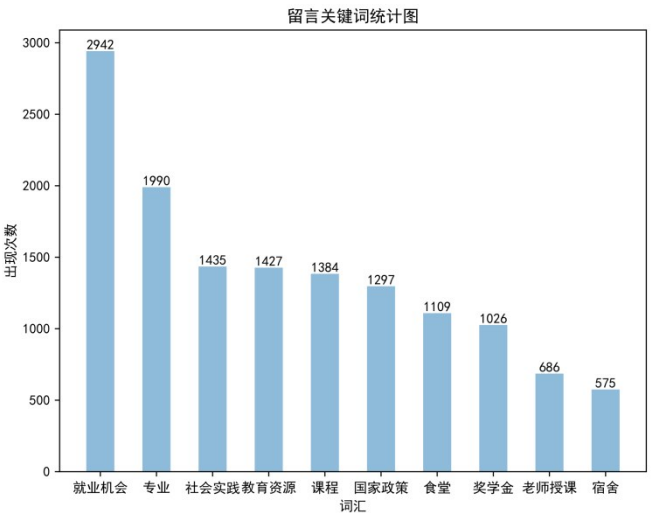


Fig. 5 Message keyword statistics
图 5 留言关键词统计

将聚类后的文本放在 Word2Vec 模型上进行训练,创建一个 Word2Vec 对象并设置参数,保存训练后的词向量。选定一个关键词作为目标词语,模型会计算目标词语与其他词语之间的相似度,并按照相似度从高到低进行排序。过滤掉一些无关词语后输出 5 个最相关的词语,输出结果如表 4 所示。可以看出,针对就业机会,学生们建议在专业教育中多一些实践训练,提高就业机会;针对专业,学生们建议降低专业调剂难度,提高专业技能;针对社会实践,学生们建议多组织课外活动,提高动手能力;针对教育资源,学生们建议加大教育投资,支持教学改革;针对课程,学生们建议合理安排选修课,多增加一些实用性课程;针对国家政策,学生们建议多出台政策激励学生学习,多资助贫困学生,增加就业岗位;针对食堂,学生们建议改善伙食水平,提高卫生标准;针对奖学金,学生们建议增加奖励名额,扩大奖励范围;针对老师授课,学生们建议老师可以改变教学方式,调动学生积极性,提高互动率;针对宿舍,学生们建议学校安装空调,改善住宿环境。

4 结语

本文利用 LDA 模型更适合处理长文本数据的优势,提出将 Mini-Batch K-Means 聚类算法加入到融合 TF-IDF 的

Table 4 Words related to target words

表4 与目标词语相关的词语

目标词语	相关词
就业机会	增加、渠道、指导、实习、创业
专业	选择、适合、职业规划、感兴趣、调剂
社会实践	课外、动手、比赛、培训、专业技能
教育资源	公开、平衡、资源分配、资金投入、公平
课程	选修课、合理安排、有趣、实用性、多样化
国家政策	出台、激励、贫困生、资金、就业
食堂	伙食、价格、卫生、降价、整改
奖学金	名额、范围、扩大、贫困生、补助
老师授课	调动、教学方式、互动、趣味性、积极性
宿舍	空调、住宿、网络、热水、改造

LDA 模型中,用于分析 CCSS 调查问卷中海量学生留言文本中蕴含的主题。相较于传统的 K-Means 聚类算法,本文算法的运行时间缩短至 10 s 以内,平均 PMI 值也有所提高,提示聚类质量有所改善。通过对 2011-2021 年 CCSS 项目的学生留言文本进行聚类分析,发现学生关注的重点为就业机会、专业、课程、社会实践、教育资源等,为学校及时发现教学管理中的薄弱环节与潜在问题,持续改进教育质量提供了重要参考。然而,本文仅通过 PMI 值来判断主题模型的优劣,评价指标过于单一。后续将尝试综合采用输出困惑度、一致性等参数全面评价模型性能,以更加精准地提取大学生留言中的重要信息。

参考文献:

- [1] GUAN H, ZONG F Y, QU P. Classification of user comments based on BTM and semantic enhancement of long text[J]. Journal of Computer Technology and Development, 2023, 33(7): 181-187.
关慧,宗福焱,曲盼. 基于 BTM 和长文本语义增强的用户评论分类[J]. 计算机技术与发展, 2023, 33(7): 181-187.
- [2] CHEN Q, ZHANG L, JIANG J, et al. A review analysis method based on support vector machine and topic model [J]. Journal of Software, 2019, 30(5): 1547-1560.
陈琪,张莉,蒋竞,等. 一种基于支持向量机和主题模型的评论分析方法[J]. 软件学报, 2019, 30(5): 1547-1560.
- [3] XIAO J M, CHEN S Z, FENG Z Y, et al. An APP maintenance and evolution method based on user review automatic analysis[J]. Journal of Computers, 2020, 43(11): 2182-2202.
肖建茂,陈世展,冯志勇,等. 一种基于用户评论自动分析的 APP 维护和演化方法[J]. 计算机学报, 2020, 43(11): 2182-2202.
- [4] ZHANG X F. Research on micro-blog hot topic discovery based on VSMS-BTM topic model [D]. Chongqing: Southwest University, 2017.
张晓飞. 基于 VSM-BTM 主题模型的微博热点话题发现研究[D]. 重庆: 西南大学, 2017.
- [5] BLEI D M, NG A Y, JORDAN M L. Latent Diriclet allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.
- [6] KANG J H, LERMAN K, PLANGPRASOPCHOK A. Analyzing microblogs with affinity propagation[C]// Proceedings of the KDD Workshop on Social Media Analytics, 2010: 67-70.
- [7] CHEN Y, SHAO X, ZHAO H B. Weibo topic mining based on UR-LDA [J]. Computer Technology and Development, 2017, 27(6): 173-177, 182.
陈阳,邵曦,赵海博. 基于 UR-LDA 的微博主题挖掘[J]. 计算机技术与发展, 2017, 27(6): 173-177, 182.
- [8] ZHAO W X, JING J, WENG J, et al. Comparing Twitter and traditional media using topic models[C]//European Conference on Advances in Information Retrieval, 2011: 338-349.
- [9] HUANG B. Research on microblog topic discovery algorithm based on vector space model and LDA model [D]. Chengdu: Southwest Jiaotong University, 2012.
黄波. 基于向量空间模型和 LDA 模型相结合的微博客话题发现算法研究[D]. 成都: 西南交通大学, 2012.
- [10] APAZA R G, CERVANTES E V, QUISPE L C, et al. Online courses recommendation based on LDA [C]// Symposium on Information Management and Big Data, 2014: 42-48.
- [11] MOODY C E. Mixing Dirichlet topic models and word embeddings to make lda2vec[DB/OL]. <https://arxiv.org/abs/1605.02019>.
- [12] QIU Y F, ZHAO B, LIN M M, et al. K-means short text clustering algorithm with semantic improvement [J]. Computer Engineering and Applications, 2016, 52(19): 78-83.
邱云飞,赵彬,林明明,等. 结合语义改进的 K-means 短文本聚类算法[J]. 计算机工程与应用, 2016, 52(19): 78-83.
- [13] ZUO J, CHEN Z M. An anomaly detection algorithm based on improved K-means clustering[J]. Computer Science, 2016, 43(8): 258-261.
左进,陈泽茂. 基于改进 K 均值聚类的异常检测算法[J]. 计算机科学, 2016, 43(8): 258-261.
- [14] SUN Z P, QIAN X Z, WU Q, et al. Adaptive rough K-means algorithm based on weighted distance computation [J]. Computer Applications and Research, 2016, 33(7): 1987-1990.
孙志鹏,钱雪忠,吴秦,等. 基于加权距离计算的自适应粗糙 K-均值算法[J]. 计算机应用研究, 2016, 33(7): 1987-1990.
- [15] HUANG J Z, NG M K, RONG H Q, et al. Automated variable weighting in k-means type clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668.
- [16] ZHOU B J, TAO Y Z, JI B. Minimizing error square sum k-means initial clustering center optimization method [J]. Computer Engineering and Applications, 2018, 54(15): 48-52.
周本金,陶以政,纪斌. 最小化误差平方和 k-means 初始聚类中心优化方法[J]. 计算机工程与应用, 2018, 54(15): 48-52.
- [17] CHEN Z Q. Research and application of improved K-Means algorithm [D]. Fuxin: Liaoning Technical University, 2023.
陈治琦. 改进 K-Means 算法的研究与应用[D]. 阜新: 辽宁工程技术大学, 2023.
- [18] LEWIS D D, SCHAPIRE R E, CALLAN J P, et al. Training algorithms for linear text classifiers [C]//International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999: 298-306.
- [19] LIANG X T, GU L. Research on Chinese word segmentation and part-of-speech tagging [J]. Computer Technology and Development, 2015, 25(2): 175-180.
梁喜涛,顾磊. 中文分词与词性标注研究[J]. 计算机技术与发展, 2015, 25(2): 175-180.
- [20] GUAN Q, DENG S H, WANG H. A comparative study of common stop word lists in Chinese text clustering [J]. Data Analysis and Knowledge Discovery, 2017, 1(3): 72-80.
官琴,邓三鸿,王昊. 中文文本聚类常用停用词表对比研究[J]. 数据分析与知识发现, 2017, 1(3): 72-80.
- [21] VELMURUGAN T, SANTHANAM T. Computational complexity between K-Means and K-Medoids clustering algorithms for normal and uniform distributions of data points [J]. Journal of Computer Science, 2010, 6(3): 363-368.

(责任编辑:尹晨茹)