Machine Translated by Google

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

社交计算高级数据挖掘方法特别章节

收到日期:2020 年 1 月 30 日,接受日期:2020 年 2 月 12 日,出版日期:2020 年 2 月 19 日,当前版本日期:2020 年 2 月 28 日。

数字对象标识符 10.1109/ACCESS.2020.2974983

开放式问题的文本挖掘
大学教师自我评价：
LDA 主题建模方法

DIEGO BUENAÑO-FERNANDEZ 1和 SERGIO LUJÁN-MORA 3 ，马里奥·冈萨雷斯1 ，大卫·吉尔2 ，

1美洲大学 (UDLA) 工程与应用科学学院,基多 170504,厄瓜多尔2阿利坎特大学计算机技术与计算系,03690 阿利坎特,西班牙3阿利坎特大学软件与计算系统系,03690 阿利坎特,西班牙

通讯作者:Diego Buenaño-Fernandez (diego.buenano@udla.edu.ec)

这项工作部分由厄瓜多尔基多美洲大学 (项目 SIS.DBF.19.02 和项目 SIS.MGR.20.02) 资助,部分由西班牙科学、创新和大学部 (项目 ECLIPSE-UA) (资助编号 RTI2018-094283-B-C32) 资助,部分由 Lucentia AGI 资助。

摘要网络上每天都会产生大量文本,包括社交网络评论、博客文章和开放式问题调查等,这表明文本数据被频繁使用,因此,处理文本数据对研究人员来说是一项挑战。主题建模是文本挖掘中的新兴技术之一;它基于潜在数据的发现和文本文档之间关系的搜索。本文的研究目标是评估一种基于主题建模和文本网络建模的通用方法,该方法允许研究人员从使用开放式问题的调查中收集有价值的信息。为了实现这一目标,我们通过使用案例研究对该方法进行了评估,该案例研究研究了厄瓜多尔一所大学的教师自我评估调查的答复。本文的主要贡献是纳入了聚类算法,以补充执行主题建模时获得的结果。所提出的方法基于四个阶段:(a) 构建文本数据库,(b) 文本挖掘和主题建模,(c) 主题网络建模和 (d) 已确定主题的相关性。在之前的研究当中,我们已经观察到人类的解释性贡献在这一过程中发挥着重要作用,尤其是在阶段 (a) 和 (d)。

因此,可视化界面（如图形和树状图)对于研究人员来说至关重要,以便主题能够有效地分析主题建模的结果。作为本案例研究的结果,本文介绍了教师在课堂上为提高学生的记忆力而采取的主要策略。此外,所提出的方法可以扩展到分析博客、社交网络、论坛等中的非结构化文本信息。

索引术语潜在狄利克雷分配、开放式问题、教师自我评估、主题建模、主题网络。

一、引言

在进行文本分析时缺乏通用的方法已成为文本挖掘领域研究的巨大挑战和空白。这是一个问题,因为每种情况使用的文本挖掘模型都不同,因为每个区域都有一组具有不同语义的特定单词。例如,用于分析社交网络 Twitter 上的消息的文本挖掘模型非常有用。

与用于分析给定调查中开放式问题答案的文本挖掘模型不同 [1]。主题建模是文本挖掘、文本中隐藏信息的恢复和社会网络分析中最强大和最广泛使用的技术之一 [2]。基于这些前提,开发一个具有通用标准的模型是适当的,该模型可以保证应用于不同领域的主题建模的有效性和可靠性。本研究强调了在任何旨在有效应用主题建模的方法中需要考虑的四个方面：(a)明确定义收集过程

副主编负责协调本稿件的审阅和批准出版的是赵永强

本作品根据知识共享署名 4.0 许可证获得许可。有关更多信息,请参阅 http://creativecommons.org/licenses/by/4.0/

Machine Translated by Google

以及文本数据库的预处理； (b)主题建模参数的正确选择； (c) 模型可靠性的评估；
(d) 对所确定主题的彻底解释。

另一方面,应该指出的是,开放式问题的答案代表了包含非常有价值信息的非结构化数据源。通过开放式问题提取信息的想法非常有吸引力,并且广泛应用于不同领域和网络平台,因为这些数据真正代表了受访者的标准[2]。

网络应用程序生成的大量文本以及在此特定情况下包含开放式问题的调查表明,文本类型数据的使用越来越多。

因此,有必要深入研究针对文本信息的自动分析方法。

自然语言处理（NLP)中使用的不同技术的应用可以管理文本文档中描述的人类语言信息 [3]。

根据文章 [4],技术发展使传统的“口口相传”传播形式成为一种新的传播形式:电子口碑 (e-WOM)。该术语由 Gold-smith 和 Horowitz [5] 创造,指的是通过上述在线应用程序和平台进行的互联网传播,这些传播会产生大量的文本信息。

主题建模侧重于对文本文档进行分组,假设每个文档都是名为主题的潜在变量的函数。在主题建模中,主题由通过适当的统计方法生成的单词列表组成 [6]。文本可以是书籍章节、博客文章、电子邮件、开放式问题的答案以及任何类型的非结构化文本。主题建模的目的不是理解文本文档中单词的语义。主题建模通常被称为“无监督”方法,因为它们涉及推理过程,而不是假设所考虑主题的内容,并且已用于社交网络、软件工程、语言科学等各种领域。[2]。

通过潜在狄利克雷分配 (LDA) 进行主题建模由 Blei、Ng 和 Jordan 于 2003 年首次提出[6],是一种计算分析技术,可用于研究文本数据集合的主题结构。该算法将归纳方法与统计测量相结合,这使得它特别适合探索性和描述性分析[7]。

在教育领域,教师评估是一个正式且系统的过程,可以衡量教师的表现。高等教育机构教学标准的制定要求教师有效地履行这些标准[8]。因此,评估教师的优点或缺点是一个至关重要的过程。高效的教师应表现出高水平的教学技能,以满足所需的责任标准,并深切关心学生及其成功。

本文提出了一个案例研究,其中研究了厄瓜多尔大学教师自我评估调查中包含的开放式问题的答案。

本大学实施的教师评估体系由多个部分组成:异质评估、共同评估和自我评估。

在教育领域,异质评价是学生对教师的评估,目的是评估他们在教学过程中的表现。大学采用的异质评价模式由五个部分组成。第一部分是指教师使用的教学方法。第二部分涵盖了学科目标的实现情况。第三部分允许学生将正在评估的课程与所修的其他科目进行比较。第四部分反映了学生对这门学科的期望。最后,第五部分允许学生就一些与教师评估相关的其他话题公开表达自己的看法。共同评价或也称为同侪评价,它是对课堂的观察,其中同一知识领域的教师作为其中一位同事的观察者。自我评估是教师对自己在教学过程中的表现进行自我分析的过程。教师是自己表现的最佳评判者,因为他们能够对自己的大部分专业发展负责[8]。

在所考虑的大学中,教师的自我评估过程是通过包含 12 个开放式问题的在线调查进行的。开放式问题提供了教师自我评估过程中使用的调查和问卷类型的重要数据。

这些数据可以为研究人员提供有关受访者态度和观点的信息,而这些信息很难从封闭式问题数据中获得 [9]。然而,开放式问题的使用也存在一系列相关的分析问题,特别是在确定与所提出的问题相符的连贯主题方面 [10]。

本研究提出应用一种方法在开放式问题的调查中执行主题建模算法,目的是提供有关教师为提高大学学生保留率而使用的策略的信息。

大多数组织现在使用在线开放式问题调查来请求利益相关者就各种主题提供反馈（例如,“我们如何改进我们的服务？”）。开放式问题成为调查的关键组成部分。它们用于识别观点并澄清研究人员以前没有想到的歧义[1]。

然而,对于大样本,手动分析这些信息几乎是不可能的。由于虚拟环境中生成了大量的文本数据,文本分析正成为一个快速发展的领域[11]。在这种情况下,主题建模技术是分析大量文本信息的有力工具。然而,缺乏一种通用方法指导主题建模的应用来评估开放式问题的答案

Machine Translated by Google

分析文献中尚未确定不同背景下的问题。因此,在本文中,我们提出应用基于主题建模和文本网络建模的通用方法,该方法允许研究人员使用开放式问题从调查中收集有价值的信息。因此,本研究涵盖了文献中确定的研究空白,

所提出方法的另一个贡献是包含文本网络建模算法,该算法与 LDA 主题建模算法的贡献一起,为所提出的案例研究提供相关结果。主题建模往往只关注术语的频率,而文本网络的分析则考虑文本的结构和所用单词的顺序[12]。

本文的结构如下:在第二部分中,我们建议回顾一些与主题模型在开放式问题分析中的应用相关的工作。第三部分描述了用于分析所考虑案例研究的文本数据的方法。

第四部分详细介绍了研究结果。
最后,第五节详细介绍了所开展工作的结论,旨在为未来的工作提供讨论点。

二.文献综述在本节中,我们描述了一
些专注于使用主题建模和文本挖掘来分析涉及开放式问题的调查的作品。

在 [11] 中,作者对软件工程领域的主题建模进行了分析,以明确主题建模在一个或多个软件存储库中的应用程度。他们重点关注了 1999 年 12 月至 2014 年 12 月期间撰写的 167 篇文章,并评估了主题建模在软件工程领域的使用情况。他们确定并展示了通过主题建模挖掘非结构化存储库的研究趋势。他们发现大多数研究仅关注有限数量的软件工程任务。

在医学领域,主题建模已在多种应用中实现。在[13]中提出了一项使用主题建模的研究,其目的是根据糖尿病患者的电子病历推断其信息需求,以推荐相关的教育材料。在本研究中,使用了机器学习语言工具包 (MALLET) 工具包,它是对自然语言处理有用的 Java 代码的集成集合。研究中提出的基于主题建模的方法可以帮助选择与特定患者相关的教育材料。使用 MALLET 对作者建立的多个主题进行主题建模;但建议以更系统的方式探讨是否有必要确定最佳主题数量。最后,研究提到,关于主题建模,
详细进行数据预处理阶段至关重要[13]。

[2] 中提出的研究提出了一种涉及结构主题建模 (STM) 的替代半自动方法。该方法结合了所考虑文件中的特定信息,例如作者的性别或政治派别。 STM 已应用于政治学等许多知识领域。本文重点分析 STM 对于处理包含大量开放式问题的调查的研究人员有何用处。这项研究使用了多项实验以及对美国国家选举研究 (ANES) 中可用开放数据的分析。研究中提出的模型允许研究人员从数据中发现主题,而不是假设它们。当确定的主题与理论观点不相符时,研究人员要么(a)修改模型以供将来使用,要么(b)保留模型并使用人工编码程序。这使得该模型可以在多个领域复制。

在[14]中,研究的重点放在文本分析的不同方法上,考虑到学术界的不同领域,例如传播研究、社会学以及在某种程度上行政和政治学。分析过程中的自动化程度与人工编码在不同领域之间可能存在很大差异,某些领域往往比其他领域更青睐某些技术。本文提供了文本分析现有技术的简化分类的定义,以及组织科学计算机辅助文本分析的具体指南的描述。

Gurgacan 等人 [15] 在软件工程领域开发的工作中,应用了一种基于 LDA 的半自动化主题建模方法。在这项工作中,一个提供完整搜索选项的在线求职网站被用作数据源。研究人员研究了网站上发布的与大数据软件工程相关的招聘广告的文本内容。

这项工作的目的是确定大数据软件工程所需的技能组合和知识领域。

这项工作的结果是开发了一种分类法,其中包括大数据软件工程所需的基本知识、技能和工具领域。

在教育领域,文本挖掘尚未得到充分利用。Erkens 等人 [16] 提出的研究建议通过开发一种自动化工具(分组和表示工具)来应用文本挖掘,该工具面向认知信息的分析和可视化,可以改善课堂上的协作学习。本文对用于开发该工具的 LDA 和向量空间模型方法进行了比较,该工具已在实验案例研究中得到验证。研究结果表明,讨论对学生学习有显著影响。此外,结果表明,当学生在讨论过程中使用开发的工具时,这种影响尤其强烈。

Machine Translated by Google

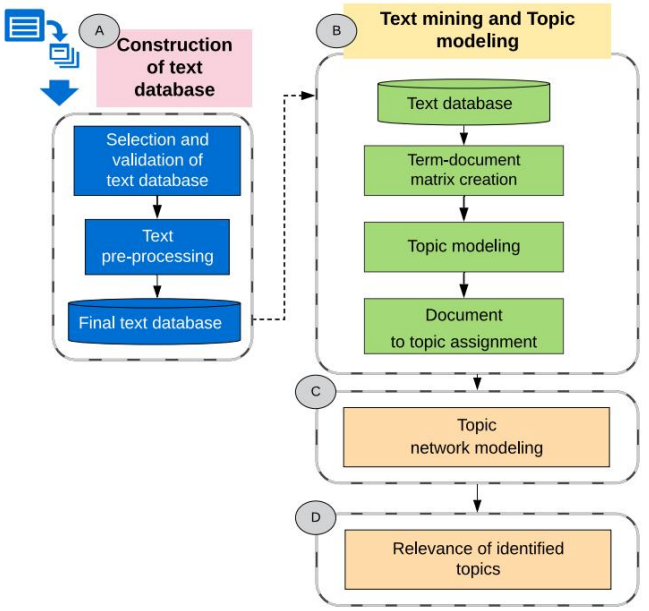

图 1.主题建模和主题网络的方法。

三.方法论本节详细描述用于将主

题建模和网络建模应用于一组非结构化文本数据的方法论。我们进行了一项案例研究,其中分析了纳入教师自我评估调查的开放式问题的答案。图1直观地描述了该方法如下:在第一阶段 (A) 中,描述了收集和验证文本数据库所需的步骤。

这意味着,一旦收集到文本信息,就需要对这些数据进行随机验证过程,以验证所收集的信息是否有助于实现研究目标。 (B) 在此阶段,首先对阶段 (A) 中收集和验证的数据执行文本挖掘过程。通过此过程,我们寻求识别文本数据库不同元素之间存在的模式和关系,以便发现否则难以识别的新信息。此外,通过应用 LDA 算法执行主题建模是对这一阶段的补充。

在该方法的第三阶段(C)中,应用了主题模型网络。该模型的目的是将任何文本表示为网络。在我们的例子中,每个主题代表一个节点,连接代表它们之间存在的关系。该模型根据术语的共现来识别文本中最有影响力的单词[17]。然后在模型中,应用数据可视化技术来识别代表文本主要主题的不同集群以及它们之间的关系。最后,在最后阶段(D),在研究领域专家的协作下,对计算机自动分组的主题进行识别和描述。应该强调的是,必须对相关主题的相关性进行分析

针对正在考虑的问题。本文强调了所确定的主题对提高大学学生保留率的策略的贡献。

A. 文本数据库的构建

在此阶段,进行主题建模过程将操作的文本数据库的收集和描述。此外,还执行文本数据库预处理。

1)文本数据库的选择和验证 在文本数据库的选择和验证方面,此阶段将进行数据源的分析,并规划收集过程。这些操作旨在构建一个可靠的语料库,以便最佳地执行文本挖掘和主题建模过程。

2)文本预处理预处理阶段是保证数据质量的重要操作,特别是在非结构化文本内容的分析中[18],[19]。在本研究中,应用于实验数据集的预处理是通过几个连续步骤确定的。所提出模型的第一步旨在消除不需要的和不相关的噪声和数据。 CSV 文件使用 UTF-8 编码过程来识别西班牙语的所有特殊字符。此外,所有回复都转换为小写,并删除了停用词、多余的空格、标点符号和数字。此外,为了丰富预处理阶段,还进行了词干提取和词形还原过程。词干提取和词形还原是单词标准化的方法。词干提取是 NLP 中使用的一种将单词缩减为词根或词干的方法。在语言学中,词形还原是将单词的不同弯曲形式分组的过程,以便可以将它们作为单个元素进行分析[20]。

对语料库中重复次数最多的单词进行同义词分析。针对这些单词,对关键词在上下文中的应用进行了研究,以确保用同义词替换其中一个单词不会改变句子的含义。此外,还将复数单词替换为单数单词,例如将“subjects”替换为“subject”。

B.文本挖掘和主题建模

1)术语-文档矩阵创建在NLP中,文档通常由词袋表示,实际上是术语-文档矩阵。单词文档矩阵是语料库的简化表示,它成为主题建模中使用的输入[21]。文档进入语料库的顺序并不意味着它们之间存在任何关系,因为在最终的术语文档矩阵中,所有文档及其术语被混合以执行

Machine Translated by Google

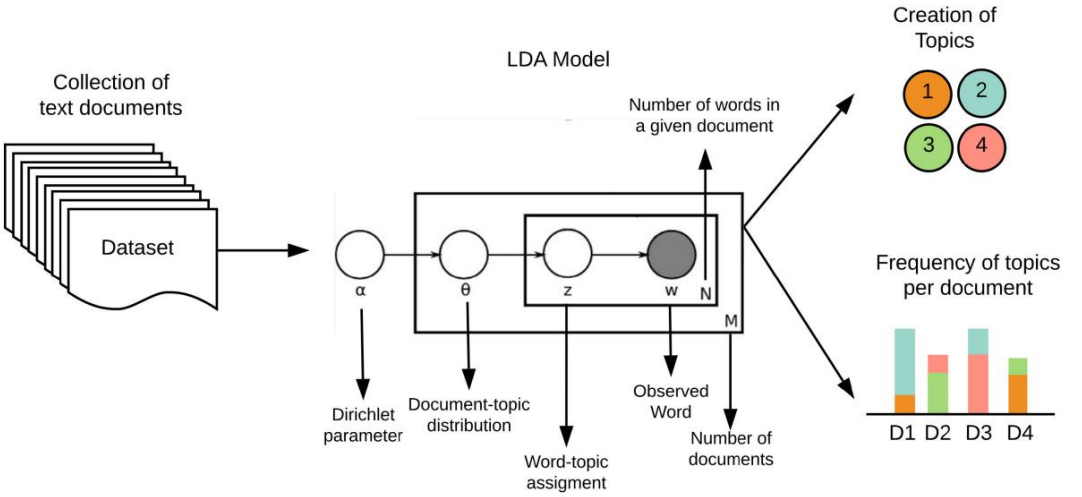

图2.LDA 算法示意图。

必要的统计分析。单词和文档的可互换性可以看作是概率潜在语义分析 (PLSA) 和 LDA 主题建模算法所基于的基本假设 [22]。

鉴于术语-文档矩阵的定义,并且取决于构成语料库的文档数量,这些矩阵往往非常大。因此,在文本挖掘中,通常会消除稀疏术语,即在极少数文档中出现的术语。

通常,通过此过程,可以大幅减小矩阵的大小,而不会丢失其固有的重要关系。

2)主题建模这是一种对文本

文档进行分组的统计方法;它基于以下前提:每个文档都是称为主题的潜在变量的函数。近年来,这种方法在计算机科学领域得到了广泛应用,特别关注文本挖掘和信息检索[1]。主题建模方法基于隐藏变量(主题)的存在,这些变量解释可观察变量(文档)之间的相似性。用于建模主题的主要且最常用的算法是LDA [6]。该算法是主题的概率生成模型,其基本思想是文档由潜在主题的随机混合组成。每个文档都被建模为词袋主题的混合,每个主题都是一个离散概率分布,定义每个单词出现在给定主题上的概率。 LDA 被认为是一种用于对语料库进行建模的无监督生成概率方法。图2详细显示了该算法的操作。 LDA 假设每个由多个单词(N) 组成的文档 (M) 可以表示为潜在主题上的狄利克雷概率分布。

其中 α 表示文档主题的狄利克雷先验权重; Z表示将单词分配给给定主题,W表示在文档 M 中观察到的单词。在本研究中,LDA 用于发现文档中出现的主题。

3) 文档到主题的分配 这也许是 LDA 算法的最大

优点。在生成过程中,算法假设一个单词属于一个主题,并且一个文档属于至少一个主题。在此前提下,需要正确选择α的参数,即每个文档的主题分布。如果选择较高的 α 值,则主题的分布将是均匀的,而较低的 α 值会阻止推理过程在某些主题中分配概率百分比。

C. 主题网络建模

将阶段 (B) 第 3 点中描述的文档到主题分配矩阵进行二值化,以便为每个文档选择最相关的主题。由此产生的将文档与主题相关的二进制矩阵可以表示为二分网络,可以对其进行投影[23]以分析主题关系。

D. 已确定主题的相关性

此阶段的目标是尝试找到一个描述主题集实质内容的标签,并由算法自动分组[24]。对由此产生的主题的解释涉及一个综合过程,其中该主题专家的贡献是一个基本要素。

四.案例研究结果

在执行图 1 中描述的建议方法之前,执行了手动语料库标记过程,其中大约 10% 的总响应被

Machine Translated by Google

表 1.手动随机分析中确定的主题和标记。

Topics	Tokens
Practical teaching	Application, practice, exercises, practical, interactive, workshops, cases, experiments, laboratory, outings, field, linking, community, visits, study, analysis, clinical.
Experiential learning	Situations, life, real, real, problems, professional, experiential, experience, personal, experiences, everyday, guests, experts, consulting.
Teaching tutorships	Customized, feedback, conversation, dialogues, leveling.
Use of technology	Video, audiovisual, technology, resources, information, classrooms, virtual, environments, platforms, digital, simulation, simulate.
Types and mechanisms of evaluation	Rubrics, evaluation, summative, systematic, periodic, continuous, permanent, tests, exams.
Group and collaborative work	Work, group, collaborative, participation, peers, couples, team.
Teaching-learning environment	Environment, treatment, pleasant, flexibility, creating, trust, atmosphere, respect, mutual, climate, cordiality, cohesion.
Personalized follow-up to the student	Follow-up, motivation, accompaniment, encourage, communication, knowledge, well-being, student, listen, empowerment, approach, personal, interest, empathic.

随机分析。该过程包括对所选答案的一般阅读,以确定教师在答案中提到的宏观主题。这一过程的结果是,初步确定了 16 个主题来总结教师的回答。在手动分析的第二阶段,一些主题被合并,留下 8 个主题。

此外,还为每个主题确定了最具代表性的标记。表1显示了在手动标记过程中确定的主题和标记。这种随机手动分析在为文本预处理过程（删除多余的空格、标点符号、停用词;词干提取和词形还原)建立标准方面提供了宝贵的信息。

将所提出的方法应用于案例研究所获得的结果详述如下。

A. 文本数据库的构建在这项研究中,我们使用了一个包含大约

900 个答案的文本数据库,这些答案对应于开放式问题：“请指出您采取了哪些策略来提高学生在课堂上的保留率而不影响学术质量。包括有关您的策略的具体示例”。

问题在于它提供了一种结构,可以直接识别教师用来提高学生保留率的策略和示例。这种结构使受访者能够以类似的方式解决问题,因此能够一致地使用词语。该问题包含在2019年3月至7月第二学期在所研究大学进行的教师自我评估调查中。教师回复的文本数据库保存在 CSV 文件中。

以下是两个老师回答的示例,以供解释：

“在我的教学经验中,我注意到,当概念和理论方面与现实生活案例相关时,学生会更加专心,因此我尝试与他们谈论与正在处理的主题相关的自己或已知的经历在班上。另一种机制是停下来讨论他们在课堂主题上的经验或标准。” “每堂课中学生必须产生的积极性至关重要。参观该行业的机构为学生提供了评估和了解在课堂上获得的知识及其对雇佣关系的重要性的机会。”

B. 使用 LDA 进行主题建模

在对语料库进行手动标记（即所用方法的第一阶段)后,执行了数据准备或语料库的预处理。从语料库中,文档术语矩阵产生了 5308 个术语和 836 个文档,稀疏度为 0.99。在删除稀疏度大于 0.99 的术语（即删除相对频率较低的术语)后,得到的文档术语矩阵有 387 个和 836 个文档,稀疏度为 0.97。进行了词频-逆文档频率 (TF-IDF) 分析,以识别语料库中的重要术语。一旦确定了这些术语,就会删除排名靠前的 TF-IDF 术语,因为它们非常突出,以至于它们出现在每个主题建模中,并且具有很高的重要性。在删除和检查探索性主题后,图3 中显示了最常见的术语。这些术语似乎均匀分布在最终建模的主题中。

Machine Translated by Google

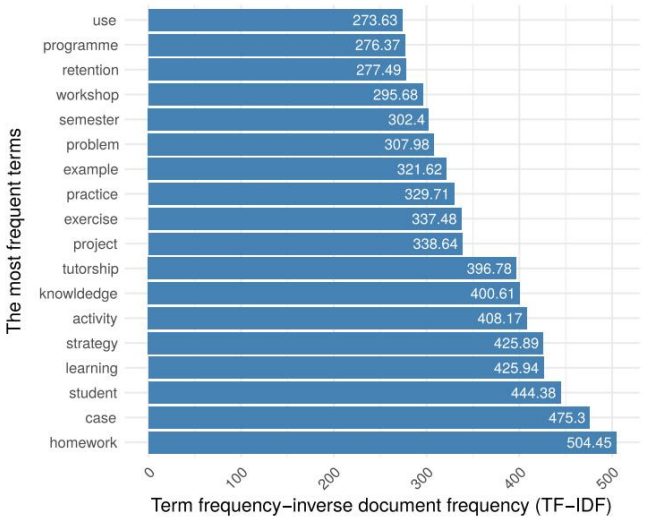

图 3.文本语料库中最常见的术语及其相应的术语频率 - 逆文档频率 (TF-IDF)。

进行探索性主题建模后,对主题数量k进行穷举搜索,并优化 LDA 算法的超参数 α。探索了从k = 2 到k = 19 的主题数量,并测试了在 0 和 1 之间均匀分布的 alpha 值,即 α = {0.01, 0.31, 0.61, 0.91}。需要评估指标来分析穷举搜索中的参数组合。为了评估参数元组(k, α),使用了一致性度量CV。CV基于滑动窗口,其中计数用于计算每个顶级单词与其他每个顶级单词的逐点互信息,从而产生一组向量,每个顶级单词一个向量。

由此,计算每个顶级词向量与所有顶级词向量之和的余弦相似度。一致性是这些相似度的算术平均值[25]。

如果所有或大多数词语（例如主题的前 N 个词语)相关,则认为该主题是连贯的。

这里的挑战是获得与手动主题标记高度相关的度量,以帮助人类进行解释[26],[27]。

图4 描述了上述穷举搜索的结果。在右上角可以看到感兴趣区域,它代表根据相干性度量的最佳参数组合。

这是对于k ⩾ 12 且 α = 0.61 和 α = 0.91 的情况。

虽然一致性的最大值出现在k = 14 和 α = 0.91 时,但在检查了这些参数的模型后,最终选择了k = 12 和 α = 0.61 的 LDA 模型,因为从人类的角度来看,主题越多,找到意义的难度就越大,而较大的 alpha 值将增加文档到主题的分配。也就是说,更高的 alpha 值将导致文档在包含的主题方面更相似 [28]。因此,我们在感兴趣的区域（右上角)选择较低的k和 α 值,以便

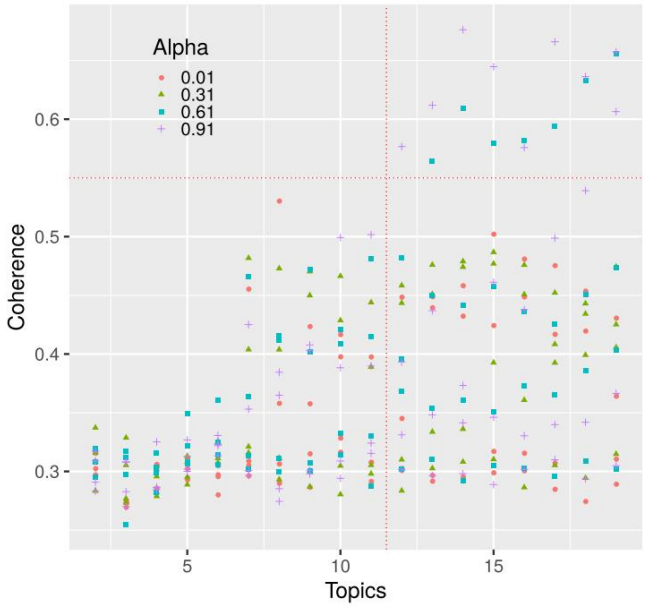

图 4.主题数量 k 和 LDA α 参数的超参数优化。

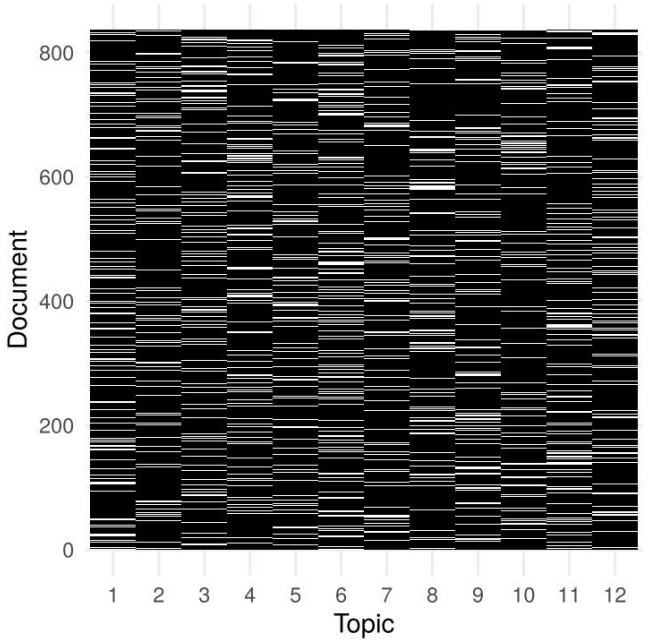

图 5.将文档（行)与主题（列)关联起来的矩阵的热图表示;白色和黑色分别对应存在连接（1)和不存在连接（0）。

具有更多的文档到主题的多样性和连接密度较低的文档主题网络。另外， k的取值与人工语料标注过程一致。

LDA 将每个文档建模为多个主题的混合。

也就是说,每个文档都有属于每个主题的概率。这可以解释为二分图,其中每个文档都连接到给定数量的主题。下一小节将探讨此类关系。

Machine Translated by Google

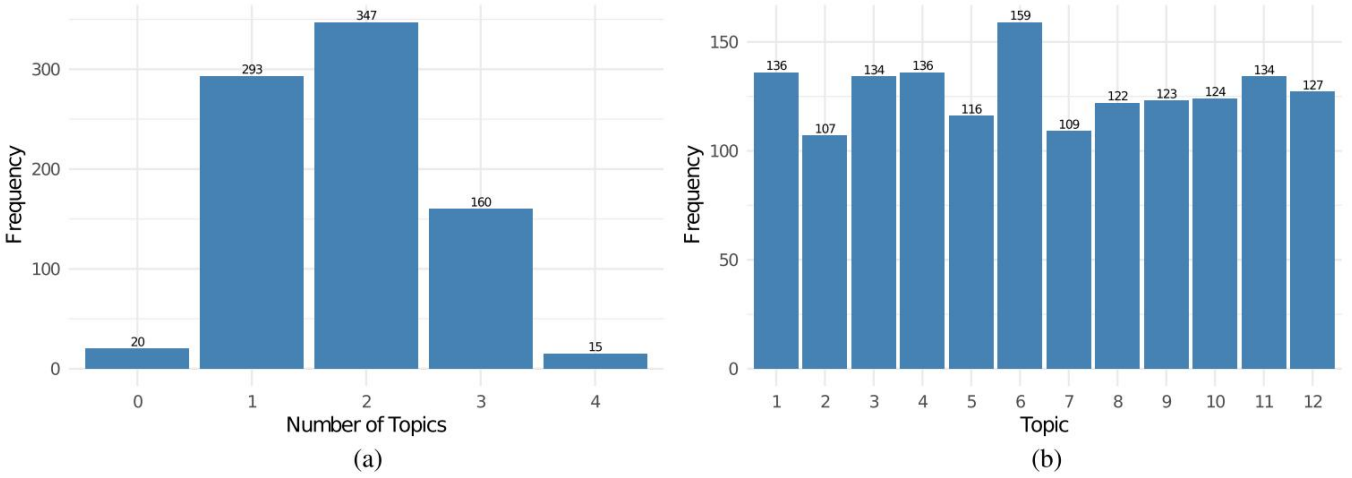

图 6. (a) 分配给文档的主题数量频率。(b) 主题分配频率。

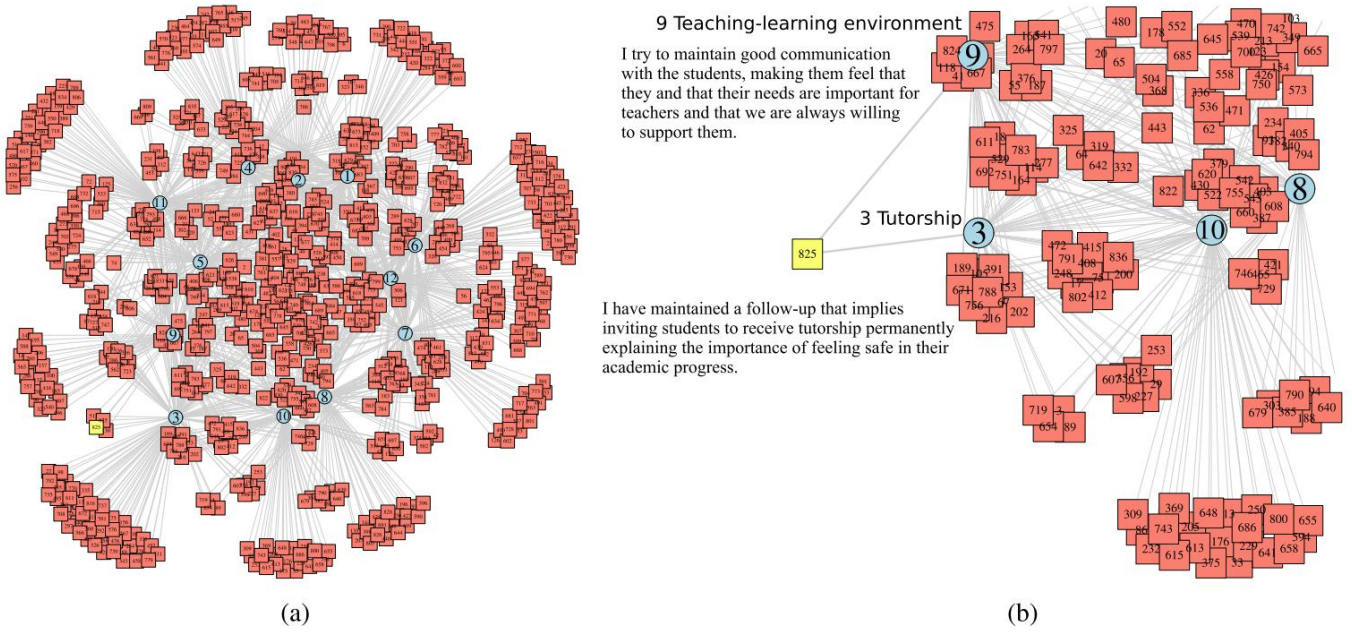

图 7. (a) 将文档（红色方块)连接到主题（蓝色圆圈)的二分网络表示。 (b) 网络的放大部分,包括教师的回答 (825) 以及与主题 3 和 9 的相互关系。

C. 主题网络建模通过使用 LDA 进行主题建模,从上一节中获得的每个

文档-主题矩阵的概率,构建一个将文档与主题关联的网络。二值化阈值 θ = 0.15 用于删除/创建文档与主题之间的关系。原始连续概率矩阵P的二值化如下:如果Pij < θ ,则B = 1,其中P是将文档与主题关联的结果二值矩阵。图5描绘了将文档（行)与主题（列)关联的每个文档-主题矩阵的概率热图,由矩阵P B 定义,其中黑色和白色分别对应 0 和 1,即存在或不存在连接。

Pij乙 = 0,否则P ij 乙

请注意,列的总和将给出主题频率,如图6 (a) 所示。分配给每个文档的主题数量如图6 (b) 所示。这是在汇总总频率后,行的总和得出的。

图7 (a) 将左图中描绘的每个主题的文档矩阵表示为二分网络,其中主题表示为淡蓝色圆圈,文档表示为淡红色方块。图7 (b) 描绘了网络的放大左下部分,人们可以在其中看到文档（红色方块)和主题（蓝色圆圈)之间发生的连接,但不能看到同一类型节点之间的连接,因为它是一个双向网络。对应于教师响应示例 (825) 的黄色方块被突出显示,

Machine Translated by Google

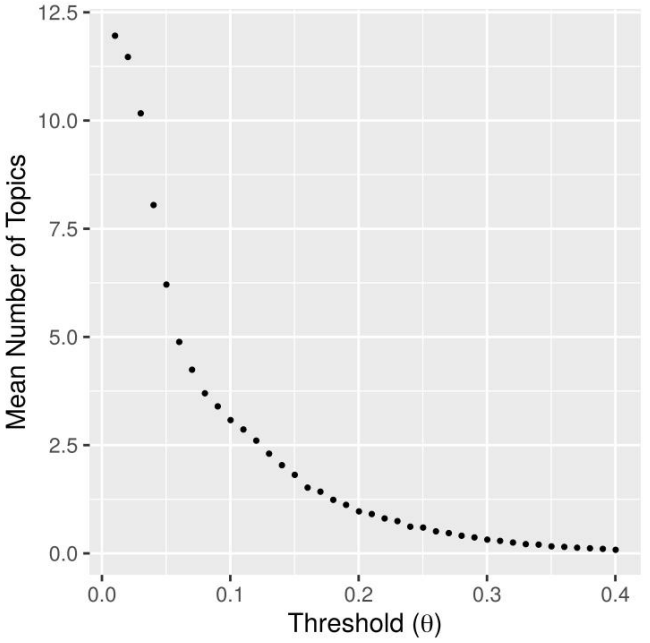

图 8.根据二值化阈值 θ 分配给文档的平均主题数。

与主题 3、辅导和相关的相关文本相关：“我一直在跟进，这意味着邀请学生接受辅导，永久解释在学业进步中感到安全的重要性。”此外，与主题的联系9、教学环境，有相关文字介绍：“我努力与学生保持良好的沟通，让他们感到他们和他们的需求对老师来说很重要，并且我们总是愿意支持他们。 ;属于同一位老师的回应（825）。

二值化阈值 θ = 0.15 是根据分配给每个文档的平均主题数经验选择的。对二值化阈值从 θ = 0.01 到 θ = 0.4 进行了详尽的搜索，并计算了每个阈值分配给文档的主题数的平均值，并将其用于确定 θ 的值。结果可以在图 8 中看到，其中肘形行为发生在 θ = 0.1 和 θ = 0.2 之间。因此，选择 θ = 0.15 的值作为上面使用的最终二值化阈值。例如，对于图6 (a)，θ = 0.15 时的平均主题数为 1.82。选择更高的 θ 值将导致网络中断开连接的文档更多（分配的主题平均数约为 0.1）。较低的 θ 值最终会导致大多数文档被分配到所有 12 个主题（分配到 12 附近的主题的平均数量），这相当于更密集连接的网络。选定的 θ = 0.15 的频率如图 6 (a) 所示。

最常见的是 1、2 和 3 个主题分配，频率分别为 293、347 和 160。只有 20 篇文档未分配任何主题（主题数量为 0）

在图7 (a) 中，有 15 篇文档被断开并从二分网络中移除，而 15 篇文档被分配到 4 个主题。
另一方面，图6 (b) 显示每个主题的频率是均匀的。也就是说，12 个主题被均匀地分配给文档。表2显示了每个主题最常见的标记。

进行二分网络投影[23]来发现所识别主题之间的关系。该投影涉及文档共有的主题，以及给出主题之间权重的共同文档的数量，从而衡量它们的关系有多强。然后，权重在 0 和 1 之间进行最小-最大缩放。主题i和j之间的距离度量dij可以写为缩放权重的补集： dij = 1 −scaled_weight。然后，考虑到上述距离，主题之间的关系在图9 （a)中被描绘为树状图。图9 (b) 将主题之间的关系描述为图表，其中社区的结构已被识别。社区揭示了网络的内部组织方式，并表明系统元素之间存在特殊关系[29]，在本例中是主题之间的关系。所选择的社区检测算法是基于边缘介数的。

这测量通过给定边的最短路径的数量。连接单独模块的边具有较高的边介数，因为从一个模块到另一个模块的所有最短路径都必须经过它们。这些边被删除以创建分层图，称为图的树状图 [30]。主题图的这种结构如图 9 (b) 所示。双边网络的另一个投影是可能的[23]，将受访者（文档）联系起来。这里没有显示这样的投影。鉴于调查的匿名性，受访者的网络中不会出现任何额外的信息以进行额外的分析。

在下一小节中，我们将讨论主题的相关性和关系。

D. 已确定主题的相关性

在主题建模过程中，方法论开始和结束时的人机交互对于确定最终模型的质量起着至关重要的作用。仅基于统计工具和计算过程的主题识别会导致文档的语义结构丢失。因此，算法识别的主题不一定描述文档的内容。在该项目开始时，对大约 10% 的教师答复进行了手动阅读。此过程允许首次识别教师在填写调查问卷时提到的主要主题。值得注意的是，将所提出的方法应用于案例研究的目的是综合教师对以下问题的回答：“指出您采用了哪些策略来提高学生在课堂上的保留率，而不会影响学术质量。包括有关您的策略的具体示例”。换句话说，

Machine Translated by Google

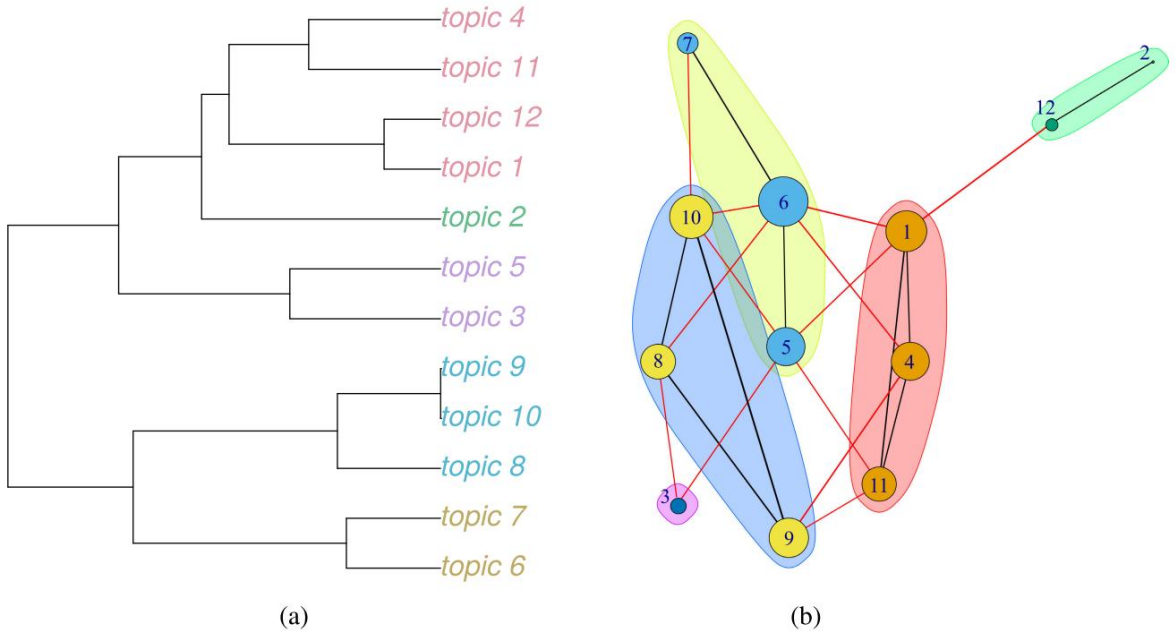

图 9.主题网络的聚类。 (a) 具有 5 个簇的树状图,使用缩放权重补集作为距离。 (b) 使用边缘介数社区检测算法的主题图和社区结构。

表 2.按主题进行的单词分类。

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	perform	learning	tracing	use	work	case	knowledge	strategy	always	professional	semester	work
2	questions	project	tutoring	virtual	case study	exercise	internship	student	have	life	final	activity
3	workshop	process	performance	tool	student	practical	application	retention	treatment	strategy	progress	tutoring
4	reading	evaluation	trouble	videos	groups	analysis	methodology	interest	importance	career	tutoring	group
5	presentation	teacher	academic	participation	different	real	makes	to get better	student	trouble	exams	group
6	content	career	tutoring	activity	same	concepts	beef up	example	teacher	student	evaluation	individual
7	readings	part	difficulties	technology	climate	examples	perform	level	activity	real	grade	practice
8	debates	teaching	tasks	through	developing	clinical	part	quality	case	personal	to get better	custom
9	discussion	feedback	evaluations	material	first	workshop	permitted	academic	trust	give	grades	student
10	base	evaluations	personal	workshop	various	games	higher	principal	they can	case	exam	following
11	participation	results	low	information	ability	study	perform	may	semester	examples	tests	account
12	investigation	elaboration	hours	active	technique	realization	theoretical	times	ambient	theory	recovery	doubts
13	related	needs	constant	communication	semester	dynamics	patients	inside	support for	know	present	teams
14	through	example	dedication	inside	method	field	simulation	question	motivation	experiences	shape	additional
15	essays	understanding	permanent	dedication	investigation	explanation	new	generate	important	I give	week	advance
16	exhibitions	virtual	student	exercise	thought	project	apply	achieve	attention	world	tasks	results
17	controls	design	tutorship	equipment	two	studies	student	face-to-face	help	chats	companions	It allows
18	then	peers	to get better	programming	foment	understand	doing	end	time	real	additional	presentation
19	discussions	autonomous	identify	support for	utilization	own	theoretical	many	start	additionally	summary	evaluate
20	information	based	custom	technique	teacher	advances	theorists	concept	keep	then	task	mistakes

本研究旨在通过半自动模型确定教师用于提高学生在大学留校率的主要策略。表 2 列出了将 LDA 算法应用于所分析的语料库后确定的案例研究的 12 个相关主题。在目前的研究中,在专家的参与下,我们致力于识别描述算法生成的每组主题的实质性内容的主题。

分配了不同的颜色来标识为表 3 中描述的每个主题贡献语义的标记。

为了更好地分析表 3 中确定和描述的主题,我们根据图9 (b) 中观察到的群集对主题进行了分组。分类如下:群集(黄色)实践学习 - 包括主题 5、6 和 7;群集(蓝色)教学主动性 - 包括主题 8、9 和 10;群集(红色)使用

表 3.按令牌组的主题标识。

Topic	Description
Topic 1	Research, analysis and reading
Topic 2	No definition
Topic 3	Tutorship
Topic 4	Use of technology
Topic 5	No definition
Topic 6	Practical learning
Topic 7	Practical learning
Topic 8	Retention strategies
Topic 9	Teaching-learning environment
Topic 10	Experiential learning strategy
Topic 11	Evaluation mechanisms
Topic 12	Team work

技术工具和传统教学策略 – 包括主题 1、4 和 11,以及集群(绿色),其中集群 12 包含与团队合作策略相关的标记。

Machine Translated by Google

另一方面,尚未确定集群 2 的具体主题。

主题 6、5 和 7 提到使用促进实践学习的策略作为教师最常用的支持学生保留的机制。在图9 (a) 和图9 (b) 中,可以观察到主题 6 和 7 的流行程度,以及它们在案例研究中的关系权重。

实践学习通过研讨会的开展、案例研究、模拟器的使用、与中心主题相关的练习、互动工作的游戏化和动态来证明。另一方面,从图9 (a)的树状图中可以看出,主题5对该主题的贡献非常轻微,这通过文本的手动分类得到了证实。

主题 10 中的术语指的是体验式学习策略,即融入专业领域的经验来激励学生学习。在图 9 (a) 的图表中,可以观察到主题 10 与主题 5、6 和 7 之间的特殊关系。这是有道理的,因为实践学习和体验式学习之间存在非常密切的关系。同一组中还有主题 8(学生学业保留)和主题 9(教学环境)。主题 9(教学环境)将与师生之间必须存在的共情关系相关的术语分组,这种关系可以激发学生的信心和动力。值得注意的是,这三个主题(8、9 和 10)与小组策略有关,这些策略不一定是课程的一部分,而是某些教师采取的特殊举措,有助于改善师生关系,从而改善教学过程和学业保留。

主题 1(阅读、分析和研究)、主题 4(技术使用)和主题 11(评估机制)属于一个新集群。主题 4 位于集群的中心,这让我们认为,在课堂上使用技术应该是吸引学生注意力的核心要素。另一方面,主题 1 和主题 11 应被视为加强虚拟环境中编程活动的工具。

值得特别注意主题 3,因为其中的术语与学术辅导的执行相关。在学术领域,这项活动对于加强学生的学习至关重要,从而提高他们在大学的持久性(保留率)。在图 5(a) 的图中,这一分析得到了证实,因为在主题 3 和主题 8(学生的学业保留率)之间观察到了直接关系。

从图9 (a)中的树状图和图表可以看出,主题 2与其他主题的关系最小,因此在研究中的相关性最小。

手动分析证实了这一点,因为该主题中包含的术语非常分散,并且不指向特定主题。

五、结论

在大学环境中,不断开展涉及教师、学生、毕业生和雇主的调查。这些调查的作用是收集有价值的信息,目的是确定上述参与者对大学学术过程的满意度。这些调查包括开放式问题,旨在发现受访者的自发想法并探索他们的态度。然而,开放式问题也面临着巨大的挑战:它们的分析需要大量的工作量和时间。这通常会避免在调查中包含开放式问题,从而失去从参与流程的人员那里收集有价值信息的机会。这就是本研究的重要性,其中我们提出应用基于主题建模和文本网络建模的通用方法,这使得研究人员能够从开放式问题的调查中收集有价值的信息。该方法的应用将允许优化对开放问题生成的文本信息进行分析所需的工作和时间。应该指出的是,所提出的方法不仅特定于教育领域,而且还可以复制到其他领域,例如本研究中描述的领域。

目前的工作从头到尾描述了通过主题建模从调查的文本数据中提取隐藏信息所涉及的方法。

本文与其他类似作品的不同之处在于应用文本网络建模作为主题建模的补充工具,以加强对开放式问题的文本分析。本文描述的主题建模案例研究旨在分析教师自我评估领域开放式问题的答案。然而,通过较小的修改,该模型就足够灵活,可以与其他非结构化数据源一起使用。

与案例研究相关的所提问题的结构提供了一个答案,该答案侧重于教师自我评估的特定要求(“...确定提高学生保留率的策略”)。这种结构允许所有教师以类似的方式回答问题,因此使用一致的词语。在计划将主题建模应用于开放式问题时,应考虑到这一基本方面。

本研究发现了一些局限性,例如数据样本相当有限。传统的主题建模方法和算法是基于文本中识别出的单词共现,因此提取短文本中的主题变得困难。然而,这种共现现象在短文本中并不常见,这意味着传统算法面临严重的数据短缺问题。

至于未来研究中要考虑的指导方针,最好使用包含更多数据的语料库。

Machine Translated by Google

因此,建议需要应用以下技术来丰富预处理阶段:消歧和词性标记、实体提取（识别专有名称,例如位置、组织或人名)和 n-gram 检测（识别分组为单个术语的单词）。另一方面,需要考虑的另一个限制是,对算法识别的主题进行的手动分析（如本案例研究中所做的那样)受到研究人员的看法和背景的影响。

在案例研究结果部分的 D 部分（已确定主题的相关性)中,在逐步应用方法中建议的阶段后,对已确定的主题进行了详细分析。在此描述中可以看出,在拟议的案例研究中,获得的主题可以清楚地识别教师为提高学生保留率而采用的主要策略。这使我们得出结论,所提出的方法可以应用于不同领域来分析具有开放式问题的调查。

根据本文中的案例研究结果,建议未来的研究应旨在确定对受访者进行分类的其他变量,例如年龄、性别、专业、临时奉献、教育水平等。其他的。这些参数将允许更好地对主题进行聚类,因此将获得更具体的结果。

参考

[1] A.-S。 Pietsch 和 S. Lessmann，“用于分析开放式调查响应的主题建模”， J. Bus。肛门,卷。 1、没有。 2,第 93–116 页,2019 年 4 月,doi： 10.1080/2573234X.2019.1590131。

[2] ME Roberts、BM Stewart、D. Tingley、C. Lucas、J. Leder-Luis、SK Gadarian、B. Albertson 和 DG Rand，《开放式调查回复的结构主题模型》,《美国政治学杂志》,第 58 卷,第 4 期,第 1064-1082 页,2014 年 3 月。

[3] H. Jelodar、Y. Wang、C. Yuan、X. Feng、X. Jiang、Y. Li 和 L. Zhao,《潜在狄利克雷分配 (LDA) 和主题建模：模型、应用、调查》,《多媒体工具应用》,第 78 卷,第 11 期,第 15169-15211 页,2018 年 11 月。

[4] A. Reyes-Menendez、JR Saura 和 JG Martinez-Navalon，“电子口碑对酒店管理声誉的影响:使用 ELM 模型探索 TripAd-visor 评论可信度”， IEEE Access,第 7 卷,第 68868-68877 页,2019 年。

[5] RE Goldsmith 和 D. Horowitz,《衡量网上征求意见的动机》， 《互动广告杂志》,第 6 卷,第 2 期,第 2-14 页,2006 年 3 月。

[6] DM Blei、AY Ng 和 MI Jordan,《潜在狄利克雷分配》,《 J. Mach. Learn. Res.》,第 3 卷,第 993-1022 页,2003 年 3 月。

[7] D. Maier、A. Waldherr、P. Miltner、G. Wiedemann、A. Niekler、A. Keinert、B. Pfetsch、G. Heyer、U. Reber、T. Häussler、H. Schmid-Petri 和 S. Adam，“在通信研究中应用 LDA 主题建模:走向有效可靠的方法”，《通信方法测量》,第 12 卷,第 2-3 期,第 93-118 页,2018 年 2 月,doi： 10.1080/19312458.2018.1430754。

[8] M. Akram 和 S. Zepeda，“教师自我评估工具的开发和验证”， Res。反思教育,卷。 9、不。 2,第 134-148 页,2015 年。

[9] JA Ross 和 CD Bruce，“教师自我评估:促进专业成长的机制”， 《教师教育》,第 23 卷,第 2 期,第 146-159 页,2007 年 4 月。

[10] WH Finch、MEH Finch、CE McIntosh 和 C. Braun，“使用主题建模与潜在狄利克雷分析和开放式调查项目”， Transl。问题心理。科学,卷。 4、没有。 4,第 403–424 页,2018 年 12 月。

[11] T.-H. Chen、SW Thomas 和 AE Hassan,《关于在挖掘软件存储库时使用主题模型的调查》,《 Empirical Softw. Eng.》,第 21 卷,第 5 期,第 1843-1919 页,2015 年 9 月。

[12] K. Lee、H. Jung 和 M. Song，“传播研究中的主体-方法主题网络分析”,科学计量学,卷。 109,没有。 3,第 1761–1787 页,2016 年 9 月。

[13] S. Kandula、D. Curtis、B. Hill 和 Q. Zeng-Treitler，“使用主题建模向糖尿病患者推荐相关教育材料”，《 Proc》。阿米亚·安努。症状， 2011,p。 674.

[14] GC Banks、HM Woznyj、RS Wesslen 和 RL Ross，“对 r（和用户友好的应用程序)中文本分析的最佳实践建议的回顾”， J. Bus。心理学,卷。 33、没有。 4,第 445–459 页,2018 年 1 月。

[15] F. Gurcan 和 NE Cagiltay，“大数据软件工程:使用基于 LDA 的主题建模分析知识领域和技能集”， IEEE Access,卷。 7,第 82541–82552 页,2019 年。

[16] M. Erkens、D. Bodemer 和 HU Hoppe，“改善课堂协作学习:基于文本挖掘的分组和表示”， Int。 J. 计算机支持的协作学习,卷。 11、没有。 4,第 387–415 页,2016 年 11 月,doi： 10.1007/s11412-016-9243-5。

[17] D. Paranyushkin，“InfraNodus:使用文本网络分析产生洞察力”,载于Proc. World Wide Web Conf.，2019 年,第 3584-3589 页。

[18] A. Kyriakopoulou 和 T. Kalamboukis，“半监督聚类对文本分类的影响”， Proc。第十七届泛希腊会议信息垫。 (PCI)， 2013 年,第 180–187 页。

[19] M. Kantardzic,数据挖掘:概念、模型、方法和算法。美国新泽西州霍博肯:Wiley,2011 年。

[20] T. Korenius、J. Laurikkala、K. Järvelin 和 M. Juhola，“芬兰文本文档聚类中的词干提取和词形还原”，Proc.第 13 届 ACM 会议信息。知道。管理。 （CIKM）， 2004 年,第 625–633 页。

[21] L. Liu、L. Tang、W. Dong、S. Yao 和 W. Zhou,《主题建模概述及其在生物信息学中的当前应用》,SpringerPlus,第 5 卷,第 1 期,2016 年 9 月。

[22] Y. Lu、Q. Mei 和 C. Zhai，“调查概率主题模型的任务表现:PLSA 和 LDA 的实证研究”， Inf. Retr.,第 14 卷,第 2 期,第 178-203 页,2010 年 8 月。

[23] DB Larremore、A. Clauset 和 AZ Jacobs,《有效推断二分网络中的社区结构》， 《物理评论 E》、《统计物理等离子体流体相关跨学科顶级期刊》,第 90 卷,第 1 期,2014 年 7 月,文章编号 012805。

[24] C. Jacobi、W. Van Atteveldt 和 K. Welbers，“使用主题建模对大量新闻文本进行定量分析”， Digit。新闻主义,卷。 4.没有。 1,第 89-106 页,2016 年。

[25] M. Röder、A. Both 和 A. Hinneburg，“探索主题连贯性测量的空间”， Proc。第八届 ACM 国际赛会议。网络搜索数据挖掘 (WSDM)， 2015 年,第 399–408 页。

[26] S. Syed 和 M. Spruit，“全文还是摘要?使用潜在狄利克雷分配检查主题连贯性分数，” Proc. IEEE 国际。会议。数据科学。Adv. Analytics (DSAA)， 2017 年 10 月,第 165-174 页。

[27] D. O Callaghan、D. Greene、J. Carthy 和 P. Cunningham，“主题建模中描述符一致性的分析”,专家系统。应用,卷。 42、没有。 13,第 5645–5657 页,2015 年 8 月。

[28] R. Deveaud、E. SanJuan 和 P. Bellot，“用于临时信息检索的准确有效的潜在概念建模”， Document numérique,卷。 17、没有。 1,第 61-84 页,2014 年 4 月。

[29] A. Lancichinetti 和 S. Fortunato,《在具有重叠社区的有向和加权图上测试社区检测算法的基准》,《 Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscipline》。Top.,第 80 卷,第 1 期,2009 年 7 月,文章编号 016118。

[30] MEJ Newman 和 M. Girvan，“查找和评估网络中的社区结构”,Phys 。修订版 E,统计数据。物理。等离子体流体相关。跨学科。顶部,卷。 69,没有。 2004 年 2 月 2 日,艺术。不。 026113。

DIEGO BUENAÑO-FERNANDEZ于 1999 年获得基多国立理工学院计算机系统工程学位,并于 2012 年获得拉丁美洲基督教大学工商管理硕士学位。他目前正在西班牙阿利坎特大学攻读计算机科学博士学位。他还是厄瓜多尔基多美洲大学工程与应用科学学院院长。他还教授操作系统和电子商务课程。他的研究方向与教育环境中的数据挖掘有关。

Machine Translated by Google

马里奥·冈萨雷斯 (MARIO GONZÁLEZ)获得博士学位。程度
2012 年获得马德里自治大学 (UAM) 计算机科学博士学
位。他专注于人工智能、复杂

系统和使用神经网络的信息处理
网络。他的研究包括建模
用于模式检索和数据的吸引子网络
分析。

大卫·吉尔 (DAVID GIL)目前是一名助理教师
计算技术系和
数据处理,阿利坎特大学。他
参与了许多国内和国际项目以及与私营公司的协议

以及与其研究相关的公共组织
他参加过许多会议,
他的大部分作品发表在国际期刊和会议上,超过

发表文章50篇。他的主要研究课题
包括人工智能应用、数据挖掘、开放数据、大数据、
以及医学和认知科学中的决策支持系统。

SERGIO LUJÁN-MORA获得博士学位。
2005 年获得西班牙阿利坎特大学软件与计算机系统系
计算机工程学士学位,

计算机科学与工程学位
1998 年毕业于阿利坎特大学。他目前
软件系高级讲师
以及阿利坎特大学的计算机系统。
近年来,他专注于电子学习,
大规模开放在线课程 (MOOC)、开放教育资源 (OER) 和
视频游戏的可访问性。他
撰写了多本书,在各种会议上发表了许多文章,
包括 ER、UML 和 DOLAP,以及高影响力期刊,包括 DKE、
JCIS、JDBM、JECR、JIS、JWE、IJEE 和 UAIS。他的主要研究兴趣
包括网络应用程序和网络开发,以及网络可访问性和
可用性。