文本分析工具——缩短从数据收集到教学干预的周期

□文 /Carolyn Rose、Chris Bogart、王旭、江师雁、李艳燕、王琦、包昊罡

Carolyn Rose

卡内基梅隆大学计算机科学学院语言技术和人机交互专业教授。着重于研究对话的社交和应用属性，并利用这种理解开发能够促进人与人、人与计算机之间对话的计算系统。领导研究小组已发表 200 多篇同行评审的研究论文。研究领域具有高度的跨学科性，包括语言技术、学习科学、认知科学、教育技术和人机交互。曾任国际学习科学学会主席、人工智能教育学会执行委员会成员。曾多次担任国际会议、研讨会和座谈会的项目联席主席。目前任《计算机支持性协作学习国际期刊》执行编辑、《IEEE 学习技术交流》副主编。

卡内基梅隆大学软件研究学院的系统科学家。研究具有跨学科性，关注开源软件、终端用户软件工程、人机交互、认知建模、神经网络算法等研究方向。研究兴趣主要在于帮助用户在团队中理解、开发和维护开源系统。获得俄勒冈州立大学计算机学院博士学位，并有超过十年的软件开发经验。

Chris Bogart

王 旭

卡内基梅隆大学计算机学院在读博士，从事人机交互和学习科学的跨学科研究。研究关注在线教育规模化、教育数据挖掘、人工智能在教育中的应用。利用学习科学理论和机器学习方法开发智能学习支持和教师支持系统，以构建更加开放、有效的在线学习平台。已发表学术文章 10 余篇，申请专利两项。

江师雁

2018 年毕业于迈阿密大学，获博士学位。现为卡内基梅隆大学博士后。主要研究方向为技术支持的 STEAM 教育、人工智能教育，以及数据可视化。

李艳燕

北京师范大学教育学部教授、博士生导师，教育信息技术学北京市重点实验室副主任。主要研究方向是计算机支持的协作学习与学习分析、教育人工智能，以及 STEM 教育。

王　琦

北京师范大学 2016 级博士生。主要研究方向为计算机教育应用，包括移动学习和泛在学习关键技术研究、适应性学习、知识图谱教育应用、人工智能教育应用。

包昊罡

北京师范大学教育学部博士生。主要研究兴趣是计算机支持的协作学习与学习分析。

如何迅速地从学习者数据中自动推断学习者的需求并提供学习支持，是学习分析领域中的一个重要研究问题。在对话分析这一子领域里，机器学习可被用于非结构化的文本数据，为教育应用提供有价值的计算模型。这一领域的大部分研究关注英文文本的分析。而本文所介绍的文本数据架构工具 DiscourseDB 和文本挖掘和建模工具 LightSIDE 也能够收集中文文本数据，以及对中文文本数据进行分析和建模，进而对中文在线讨论学习提供实时的支持。

一、背景介绍

在学习分析领域中一个重要的研究问题是如何迅速地从学习者数据中自动推断学习者的需求并提供学习支持。近期，学习分析和教育数据挖掘领域发展迅速，涌现出了一些成功案例，将机器学习应用于教育数据，以有效地揭示学生的学习过程，进而为学生提供实时的智能支持。当然，这些技术仍然处于发展的过程中。现在机器学习技术的发展为教育研究者提供了一个很好的契机，来思考机器学习模型如何为大量的学习者提供个性化的学习支持。

我们在本研究中关注基于讨论的学习。在对话分析这一子领域里，很多研究（如 [1] [2]）表明，机器学习可被用于非结构化的文本数据，为教育应用提供有价值的计算模型。在这一子领域中，如何从文本数据中获得高效的、可解释的、有应用价值的计算模型是一个重要研究目标。除此之外，搭建数据框架和分析流程来存储和处理这些文本数据也是十分有挑战性并且必要的 [3]。在这一领域的大部分文献都关注英文文本的分析。在本研究中，我们将介绍一个处理中文文本的数据框架来解决上述问题。本文介绍的工具都是开源的，可以在 LearnSphere[4] 和 DANCE[5] 项目的网站上找到。

在本文中，我们将介绍一系列文本分析工具，帮助研究者处理和分析中文文本数据。我们将讨论北京师范大学对这些工具的使用案例。最后，我们总结案例中出现的问题并展望后续研究。

二、从数据收集到系统开发的周期

最近，监督式的机器学习模型被用来对学生学习过程中的讨论进行自动评分，也就是合作学习过程的自动分析。合作学习过程的自动分析的价值体现在几个方面：1. 可以实时地监测学生在合作学习过程中的学习状态；2. 在合作学习课堂中，通过自动分析引发一系列的智能干预；3. 促进大规模、多课堂的合作学习分析和评价。这种实时动态的分析、支持方式已在研究中被证明比静态的干预方式更有效 [6]。早期对合作学习过程的自动分析关注于文本交互和点击数据 [7]，对于语音数据的分析也在逐渐发展 [8]。一个值得关注的研究发现是，由语言学或心理学理论驱动的分析模型是最有效的 [9][10]。

[1] Buckingham-Shum, S. et al.(2013). DCLA13: 1 st International Workshop on Discourse-Centric Learning Analytics.
[2] Rosé, C. P. (2018). Discourse Analytics.
[3] Jo, Y. et al. (2016). Expediting Support for Social Learning with Behavior Modeling.
[4] https://pslcdatashop.web.cmu.edu/LearnSphere
[5] http://www.cs.cmu.edu/~dance/
[6] Kumar, R. et al. (2007). Tutorial dialogue as adaptive collaborative learning support.
[7] Rosé, C. et al. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning.
[8] Gweon, G. et al. (2013). Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation.
[9] Rosé, C., & VanLehn, K. (2005). An evaluation of a hybrid language understanding approach for robust selection of tutoring goals.
[10] Wang, X. et al. (2016). Towards triggering higher-order thinking behaviors in MOOCs.

与此同时，搭建储存文本数据的框架以方便后续的数据分析是非常重要的。对数据进行计算性分析的基石是数据的表征。我们过去在研究中所使用的数据包括文本聊天数据 [11] 以及面对面会话数据的转录文本 [12]。这些数据可以通过一种统一的形式来表征。然而，当需要使用的数据包括 MOOC 或其他在线社区中的讨论帖时，文本数据的存储形式就会更加多样，因为不同平台所记录的文本数据格式是不同的。我的研究提供一个公开的文本数据存储工具 DiscorseDB（见图 1），将不同平台来源的数据以相同的表征方式存储。文本数据以 Discourse（文本集合）的形式存储。Discourse（文本集合）中包含 Contributions（每一条贡献）。每条文本之间可能有 Relations（联系）。每一条贡献又包含了 Content（内容）和 Annotations（标注）。这个对不同平台数据的统一表征，为后续应用计算模型分析文本数据提供了基础。最初版本的 DiscourseDB 只应用于英文数据。最近，DiscourseDB 被拓展以处理其他语言的数据，比如中文。

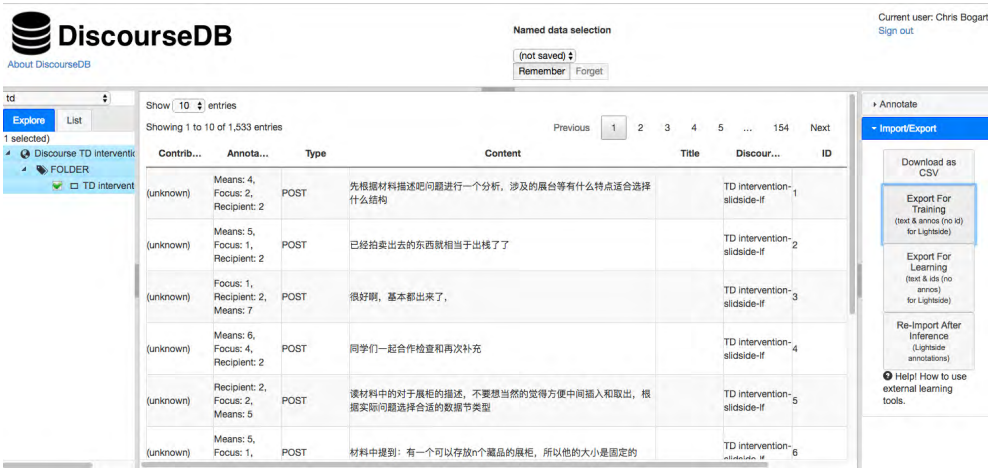

图 1：中文版本的 DiscourseDB。这个 SQL 数据库可以通过统一的格式表征不同来源的文本数据。

对合作学习过程的自动分析需要通过有监督的机器学习来实现。LightSIDE 是可以对合作学习过程进行自动分析的工具，其图形化界面如图 2 所示。LightSIDE 将文本数

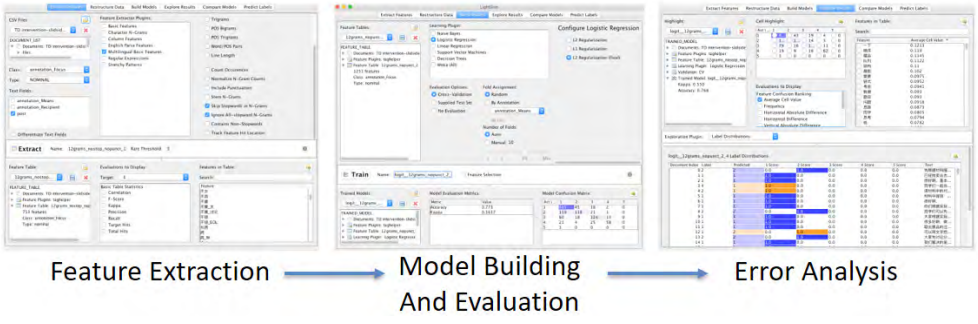

图 2：LightSIDE 的主要用户界面。它支持对中文文本数据进行迭代性建模。主要使用流程包括三步：提取文本特征；构建模型；错误分析（用于迭代性地调整模型）。

[11] Howley, I. et al. (2013). Linguistic Analysis Methods for Studying Small Groups.
[12] Clarke, S. et al. (2013). The Impact of CSCL Beyond the Online Environment.

据结构化，使自动化过程分析成为可能。LightSIDE 帮助用户从文本数据中抽取结构化特征，如单词特征，然后基于这些特征构建计算模型。其获得的计算模型可以被用来预测新的数据实例。LightSIDE 同时支持迭代式的错误分析，用来不断改进和优化计算模型。

对文本数据的表征将在很大程度上决定后续模型的优劣。文本数据的表征过程可以被想象为：有一个特征提取器在持续向文本提问，文本给出的答案即是这个特征相应的值。想象一下：一个人是由 20 个问题来定义的。现在你的任务是根据这个人在 20 个问题上的答案将他的社会类别进行分类。如果 20 个问题是经过仔细构建的，那么你可以做出准确的预测。但是，我们必须承认，在此过程中，仅仅由 20 个问题来定义一个人，将会丢失关于这个人的大量信息和见解。一旦重要信息在表征过程中丢失，无论后续算法多么先进，都无法准确地完成分类。由此可见，对文本数据的结构化表征在模型训练中起着至关重要的作用。

文本挖掘问题中使用的最典型的特征提取器就是所谓的单词特征。在提取单词特征时，对于文本中出现的每一个词都存在相应的特征在表征是否存在这个词。虽然单词特征对于训练机器学习模型通常有较好的效果，但是模型通常不能概括到其他情形。

如果追溯自然语言数据作为自动分析目标的几个领域的历史，我们会有类似的发现，即有效建模的关键是有意义的特征设计和提取。例如，最早的一个例子是自动化论文评分 [13]。最早的方法使用简单模型（如回归分析）并使用基本特征（如计算平均句子长度、长词数和文章长度）。这些方法可以可靠地给出分数，但是它们因使用的评估证据缺乏有效性而受到批评。在后来的研究中，研究者尝试提取类似于教师手动评分时会观察的特征。这包含了以内容为中心的特征，包括类似于因子分析的技术（如潜在语义分析 [14] 或话题模型 [15]），以实现基于内容的评估。其他因子分析语言分析方法，例如 CohMetrix[16] 最近被用于评估学生在多个方面的写作，包括认知复杂性等因素。在科学教育方面，LightSIDE[17] 是一项免费提供的软件工具套件，支持非专家使用文本挖掘技术，帮助教师成功评估开放性题目。LightSIDE 中包含各种基于词汇、句法和模板的特征提取器。

三、LightSIDE 中文版本

在本文中我们介绍 LightSIDE 中文版本的开发研究。英文版本的 LightSIDE 首先将英文文本分成单词（使用空格和标点符号），然后用户可选择性地应用其他分析（如使用词性词典标记每个单词）删除停用词（仅具有语法的单词，如“the”和“of”这样的单词），或者将词转化为词干（用一个共同的词干替换一个词的变体，如将“run”、“running”

[13] Shermis, M. D., & Burstein, J. C. (2003). Automated essay scoring: A cross-disciplinary perspective.
[14] Foltz, P. W. (1996). Latent semantic analysis for text-based research.
[15] Blei, D. M. et al. (2003). Latent dirichlet allocation.
[16] McNamara, D. S. et al. (2014). Automated evaluation of text and discourse with Coh-Metrix.
[17] Mayfield, E., & Rosé, C. P. (2013). LightSIDE: Open source machine learning for text.

和“ran”改为“run”）。尽管 LightSIDE 已经支持中文字符集，但为实现中文分词，我们使用了斯坦福大学开发的中文分词库 [18]。该库试图将几个字符的每个序列与字典中的单词进行匹配，并使用 CRF（条件随机字段）技术找到最好将文本分解为已知单词而没有剩余字符的分段。在以这种方式分割之后，LightSIDE 提供中文停用词、标点符号和中文词性词典的列表，可以启用其他特征提取器。

四、案例分析

在本节中，我们描述两个案例，说明如何使用 LightSIDE 来构建中文文本分析模型。第一个案例是提供自动反馈以提高讨论论坛中的帖子质量，第二个案例是分析聊天室中的互动。本节最后讨论了在不同环境中使用 LightSIDE。

案例 1：大规模在线课程中用以提升讨论质量的讨论帖的自动反馈功能

在大规模在线开放课程（MOOC）中，教师通常没有足够的时间为每个学生论坛中的发帖提供实时反馈。在本研究中，我们运用 LightSIDE 设计和开发了在线学习评论自动反馈系统，在一门 MOOC 中为学生的发帖提供实时反馈。该系统的设计和开发包含以下四个阶段：

阶段 1：问题确认

按照前文所述，课程的评论和讨论是反映学习者对课程内容理解的重要方面。而在之前的课程中，我们发现，在线学习者的评论和发帖相对较短、思考的层次较低。这种交流导致大量用户慢慢不再发帖甚至从课程中退出。为了解决这个问题，我们希望通过 LightSIDE 来为学生提供自动反馈。

阶段 2：问题解决的理论框架选取

在有监督的机器学习中，我们首先需要建立一套表征在线评论质量的理论框架。如果没有理论框架，我们就很难判断两个评论中哪个质量更高。例如，A（“是否有人想要在北京建立一个研究团队”）和 B（“基于我在 Scratch 方面的经验，我们需要建立一个研究团队探究通过角色的会话模拟是否能够促进初学者的活动参与”）。可以发现，A 只是在尝试性地提出想法，而 B 在提出问题的同时通过举例论证提出了解决方案。因此，对不同的评论提供有意义的反馈，需要一种能够对它们进行区分的框架。

本案例中，我们采用了知识建构理论来衡量学习者在在线评论中的发帖层次。一旦学习者的发帖层次得到确定，就可以为其提供适应性的引导和建议，进而促进学习者学习效果的改进。通过调研，我们发现，Gunawardena 等人 [19] 提出的分类框架能够很好地实现评论问题层次的分类，主要包括：（1）P1 分享观点（如“对焦是摄影中很

[18] https://nlp.stanford.edu/software/segmenter.html

[19] Gunawardena, C. N. et al. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing.

重要的概念”）；（2）P2 提出领域中的问题（如“如何判断我拍摄的照片对焦的好坏呢，有哪些技巧？”）；（3）P3 讨论交流观点（如“有时侯我们使用人工对焦的方法回避自动对焦更好，因为它会让你对对焦的手法更加熟悉”）；（4）P4 评价存在的观点和解决方案（如“xx 同学提交的作品很好地应用了对焦的手法，背景的虚化适当，但在 xx 方面需要提升”）；（5）P5 深层次反思（如“通过摄影课程，学习者不仅可以习得多种摄影的技术和技巧，如可以使用对焦的手法变换控制前景和背景来提升视觉效果，还能够提升学习者的审美水平”）；（6）P6 与所学内容基本无关的回复（如“这门课很不错”）。

阶段 3：基于选定的框架采集和标注数据，用以训练模型

在本阶段中，我们从三门教育技术领域的在线课程中采集了 1352 条评论数据，并按照上一阶段选取的编码进行标注。对训练数据集编码完成后，研究者通过 LightSIDE 训练计算模型。研究者在 LightSIDE 中采用了 N-gram 进行特征抽取，并采用逻辑回归对知识建构的层次进行拟合。最终模型准确率为 0.756，Kappa 为 0.659。

阶段 4：将模型整合到在线学习平台中支持即时反馈

在课程的实施过程中，我们将模型嵌入到讨论区并对讨论区进行自动反馈。当学习者在论坛提交评论时，服务器端即可接受评论内容，并通过训练好的机器模型对其进行自动分类。服务器端返回分类结果（P1-P6），并根据分类结果为学习者实时提供相应的引导和提示。例如，若用户评论被分类为 P6，说明评论质量较低，因此系统会为其提供“您的评论内容有待提升或者偏离主题，继续学习内容再来参与讨论吧”这样的反馈。又例如，若用户评论为 P5，系统会为其提供“您对知识内容的反思非常深刻，要继续保持哦”这样的反馈。在讨论区的页面呈现中，所有学习者的评论都会被打上相应的标签，方便其他学习者对不同层次的评论进行分类浏览。

本研究中，该模型已经应用于一门 1000 人的在线课程（www.etc.edu.cn）。结果表明，我们设计的自动反馈系统能够促进学习者在高层次评论数量的提升，有效降低学习者的无意义评论，这对于提升学习者的学习效果至关重要。在未来的分析中，我们还需要结合学习者的在线学习行为提供引导和反馈。

案例 2 在线讨论中基于交互的社会情感分析

在线协作学习中的情感交互对于提升小组学习动机，促进高质量的协作具有积极的作用。我们利用 LightSIDE 训练了计算模型以自动化的分析情感交互。

第一步：选择社会情感交互的分析框架

在这一阶段，我们基于 Bakhtiar 等人 [20] 提出的社会情感交互分析框架，设计了包括积极情感、消极情感、表达和其他四个维度的情感分析框架。其中，积极情感是指道

[20] Bakhtiar, A. et al. (2018). Regulation and socio-emotional interactions in a positive and a negative group climate.

歉、幽默、鼓励同伴加入讨论，以及建立信任的相关话语。消极情感包括阻止他人参与讨论，强调个人需求，以及给予他人压力等行为。表达包括各种表情包以及与情感相关的话语。其他无关的话语被标记为其他。

第二步：收集并标注训练模型

我们在某综合性大学主办的英文协作阅读活动中收集了相关的聊天信息。在活动中，学生四人一组，每组学生均采用 QQ 软件建立了自己的聊天群进行活动的讨论。在本次活动中，我们共收集了超过 14000 条聊天记录。我们随机选取 2368 条聊天日志作为模型训练的数据，由研究人员手动标注，进而用 LightSIDE 训练模型。最终模型的准确率为 0.67, Kappa 为 0.67。

采用 LightSIDE 进行自动化的分析社会情感交互是一次创新的尝试。在本次实验的基础上，可以进一步设计智能对话机器人来促进学生的社会情感交互。例如，当一个小组成员的想法不断被忽视，产生负面情绪时，智能对话机器人可以让小组更加关注被忽视的学生。

五、总结与展望

在本文中，我们介绍了一套工具，包括文本数据架构工具 DiscourseDB 和文本挖掘和建模工具 LightSIDE。这一套工具可被用来收集文本数据，以及对文本数据进行分析和建模，进而对中文在线讨论学习提供实时的支持。我们在本文中描述了这一套工具的操作方法，并通过两个早期案例表明其对中文讨论式学习进行自动化支持的可操作性。接下来，我们需要进一步推广这些工具，使我们研究团队以外的其他教师或研究者能够使用这些工具。我们提供这些工具作为资源，希望它们可以在中文语言范围内的学习、科学研究中被应用、推广，并产生积极的影响。

查看内容精选