以用户为中心的交互式人在回路主题建模系统

郑方[1], Lama Alqazlan[1],刘都[2],Yulan He[1,3,4] 和 Rob Procter[1,41]英国华威大学电脑科学系[2]英国伦敦大学城市学院贝叶斯商学院[3]英国伦敦国王学院信息学系[4]英国艾伦·图灵研究所 {Z.Fang.4, Lama.Alqazlan, Rob.Procter}@warwick.ac.uk yulan.he@kcl.ac.uk, Du.Liu@city.ac.uk

arXiv:2304.01774v1 [cs.CL] 4 Apr 2023

摘要

人在回路主题建模将用户的知识融入建模过程，使用户能够迭代式地优化模型。近期研究已证明用户反馈的价值，但仍有一些问题需要考虑，例如难以追踪变更、比较不同模型，以及缺乏基于真实世界使用案例的评估。我们开发了一种新颖的交互式人在回路主题建模系统，该系统具有用户友好的界面，使用户能够比较并记录自己采取的每一步操作，同时还提供了一种新颖的主题词建议功能，以帮助用户提供忠实于真实值的反馈。我们的系统不仅支持传统主题模型所具备的功能，即从整个语料库中学习主题，还支持目标导向的主题建模，即针对语料库的特定方面学习主题。本文概述了该系统，并介绍了一系列用户研究的结果。这些研究旨在评估该系统在逐步趋于真实的主题建模应用场景中的价值。

1 引言

每天都会产生海量非结构化文本数据。随着可用数据不断增多，要从中搜索、理解并发现知识也变得愈发困难。由于需要投入大量人工，传统的定性方法（如扎根理论）（Glaser 等人，1968）已不再适用于如此庞大的数据量。近年来，主题建模作为一种潜在解决方案，受到越来越多的关注（Heidenreich 等人，2019年；Curiskis 等人，2020年；Dantu 等人，2021年；Goyal 和 Howlett，2021年），可帮助用户组织、搜索和理解大量信息。主题建模是一种无监督机器学习技术，用于识别大型非结构化文本语料库中的隐藏主题，其中每个隐藏主题由一组描述共同主题的词语表示。用户可以轻松识别每份文档中的主题，并搜索与特定主题密切相关的文档，以开展更深入的研究。然而，传统主题模型生成的主题往往缺乏连贯性，且包含许多无关词语（Chang 等人，2009年；Mimno 等人，2011；Boyd-Graber 等人，2014年）。尽管可以通过预处理目标数据来源来解决这些问题，例如从词汇表中删除无关词语，并调整模型的超参数（如主题数量），但这要求使用者熟悉该算法。因此，对于不了解主题建模的人来说，这并不容易。

人在回路主题建模（HL-TM）将人类知识融入主题建模过程，以解决上述问题。它允许并非主题建模专家的用户通过一系列细化操作来改进模型，例如向主题添加词语、从主题中删除词语、拆分主题或合并主题（Jagarlamudi 等人，2012；Wang 等人，2012；Choo 等人，2013年；Hoque 和 Carenini，2015年；Lund 等人，2017年；Smith 等人，2018年）。虽然这些研究大多没有将细化操作纳入迭代式重新训练过程，Smith 等人（2018年）实现了一个完全交互式、以用户为中心的 HL-TM 系统，并研究了不可预测性、信任和缺乏控制等交互式系统中出现的问题如何影响用户体验。然而，他们的工作仍存在一些局限。首先，他们的系统只允许用户按顺序细化模型，也就是说，用户更新模型后，新模型会覆盖旧模型。这使用户无法比较对同一模型应用不同细化操作所产生的效果，因此难以找到最合适的操作。此外，由于旧模型无法再访问，用户可能难以判断新版本是否确实有所改进。

其次，他们的系统不允许用户在底层模型更新后撤销所做的更改。如果用户做出不恰当的更改并导致意外结果，这就会带来问题。尤其令人沮丧的是，用户花费大量时间细化模型后，整个效果却可能因一次不恰当的更改而毁于一旦。第三，他们使用潜在狄利克雷分配（LDA）（Blei 等人，2003年）作为底层模型。

由于潜在狄利克雷分配是一种全分析模型，只能从整个语料库中学习主题，因此它的应用仅限于用户对语料库中的主题没有任何先验假设的情况。用户可以通过添加先验知识来细化模型，试图将一个无关主题转变为关注目标方面的主题。然而，潜在狄利克雷分配中的高阶共现现象（Heinrich，2009年）可能会导致与目标方面相关的低频词无法被分配到无关主题中。例如，面对一组关于健康的帖子，研究人员可能特别希望分析食物对健康的影响。他们会将“食物”“吃”等食物相关词语添加到一个无关主题中。如果这些词语在帖子中的出现频率相对较低，系统就可能无法将该无关主题转变为与食物相关的主题。第四，一个尚未得到探讨的重要问题是，如何在用户的假设与现实不符时向用户发出提示（Kumar 等人，2019年），因为这可能会使细化过程产生偏差。例如，用户可能认为科技相关主题应包含“apple、google”这样的词语，而输入语料库中的“apple”却指水果。将“apple”添加到该主题后，主题就会混入与水果相关的词语。这种不恰当的细化会导致结果不佳。

为解决这些问题，我们实现了一种新颖的 HL-TM 系统，支持六种细化操作，包括添加词语、移除词语、交换词序、移除文档、合并主题和拆分主题 (Lund 等人，2017年；Smith 等人，2018年）。与Smith 等人(2018年) 的方法不同，在该方法中，用户只能按顺序细化模型，且无法撤销所做的更改；在我们的系统中，用户可以对同一模型节点进行不同的细化尝试，并比较由此产生的模型。系统还会展示完整的细化历史记录，让用户能够从第一步开始跟踪所做的更改。这也确保用户在应用了不合适的细化操作时，可以恢复到先前的模型节点。我们的系统没有采用潜在狄利克雷分配（LDA）作为底层模型，而是采用了（Fang 等人，2021年）提出的查询驱动主题模型（QD-TM）作为底层模型。QD-TM 的优势在于，它不仅支持潜在狄利克雷分配的全分析能力，即从整个语料库中学习总体主题，还支持用户进行定向分析，即通过减轻潜在狄利克雷分配中高阶共现现象的影响，学习聚焦于语料库特定方面的主题（Fang 等人，2021年）。

此外，我们还实现了一项新颖的自动主题词建议功能，引导用户为选定主题添加合适的词语。该功能会提取与主题相关的候选词列表，供用户选择并添加到主题中。评估结果表明，该功能确实有用：建议的词语与选定主题密切相关，并且与真实值更加吻合。我们还开展了一系列用户研究，旨在评估该系统在现实世界应用场景中的价值。

本研究作出了以下贡献：(1) 开发了一种新颖的交互式 HL-TM 系统，配备先进的用户界面，允许用户从同一模型节点训练不同的模型，并撤销已应用的不恰当更改；(2) 使用 QD-TM 作为底层模型，以支持全分析和定向分析主题建模功能；(3) 开发了一种新颖且高效的主题词建议功能，引导用户为选定主题添加合适的词语；(4) 开展了一项小规模用户研究和两项详细研究，旨在复现现实世界的应用场景。

2 背景

查询驱动主题模型 是一种由 Fang 等人(2021年) 开发的半监督主题建模算法。它允许用户通过单词或短语指定一个简单的查询，并返回与查询相关的主题。原始模型包含一个两阶段建模过程；在第一阶段，模型为每个概念推断一个主题，同时推断其他不相关的主题；在第二阶段，模型将每个概念的主题扩展为一组子主题。在我们的工作中，我们只关注模型的第一阶段。如果没有输入查询，该模型就会作为传统主题模型运行。

该模型基于一种层次

狄利克雷过程 (HDP) (Teh 等人，2006)，这是一种非参数贝叶斯模型，假设一家餐馆（即一个文档）拥有一组桌子，并从全球菜单中供应菜肴（即主题）。对于所有坐在同一桌的顾客（即词语），一道菜肴只会在一张桌子上供应。我们基于吉布斯采样技术开发了该算法的变体。对于文档 w_{ji} 中位置 j 的单词 i，对现有桌子 t 进行采样的概率为：

$$p(t_{ji}=t \mid t^{-ji},k) \propto \mathbb{1}(w_{ji},k_{jt})n_{jt}^{-ji}f_{k_{jt}}^{-w_{ji}}(w_{ji}) \quad (1)$$

以及抽取一个新表格 t^{new}的概率为：

$$p(t_{ji}=t^{new} \mid t^{-ji},k) \propto \alpha p(w_{ji} \mid t^{-ji},t^{new},\mathbf{k}) \quad (2)$$

此处， t^{-ji} 是所有其他单词的桌子分配。 k_{jt} 是文档 j中桌子 t 的主题分配。 n_{jt}^{-ji} 是文档j 中桌子 t 上的单词数， $f_{k_{jt}}^{-w_{ji}}(w_{ji})$ 是w_{ji} 被分配到主题 k_{jt}的概率。$\mathbb{1}(w_{ji},k_{jt})$ 是一个指示函数；如果w_{ji} 是与主题 z 相关联的预定义单词，则取值为 0，否则取值为 1。 α 是模型的超参数。为新桌子抽取主题 $k_{jt^{new}}$的概率为：

$$p(k_{jt^{new}} \mid t,k^{-jt^{new}}) \propto \mathbb{1}(w_{ji},k_{jt})m_k f_k^{-w_{ji}}(w_{ji}) \quad (3)$$

人在回路主题建模近年来受到了广泛关注。Boyer 等人（2017年）创建了一种人机协作方法，用于在大型文本语料库中识别系统思维主题。该方法要求用户主观地识别并提供描述该主题的种子文档，以指导主题模型的训练。然而，这种方法没有将任何先验知识纳入建模过程，而只是对训练语料库进行了修改。为了更好地利用人类知识进行主题建模，实现了多种细化操作，例如添加词语、移除词语、添加文档、移除文档、创建主题、合并主题或移除主题（Hoque 和 Carenini，2015年；Lund 等人，2017年；Smith 等人，2018年）。在模型采样过程中，这些细化操作会改变底层模型的先验知识。为了了解不同细化操作的实用性，Lee 等人（2017年）采用以用户为中心的方法，探究非专家用户如何解读主题模型，以及他们最希望使用哪些细化操作，但他们只实现了一个不具备完整交互功能的基础系统，因此所应用的细化操作并未更新底层主题模型。

Smith 等人(2018年) 在 Lee 的工作基础上向前迈进了一步，实现了一个完全交互式、以用户为中心的 HL-TM 系统，并进一步研究了不可预测性、延迟和信任等常见的交互式机器学习挑战如何影响用户体验。Kumar 等人(2019年) 首次对不同的细化操作以及两种反馈注入框架进行了比较评估，即知情先验和约束。他们不仅指出，未来的研究应使用最终用户测试系统，因为他们的实验仅采用了模拟用户行为，还强调了当用户的假设与现实不符时，向用户发出提示的重要性。尽管还存在其他 HL-TM 系统，例如 UTOPIAN（Choo 等人，2013年）和 ConVisIT（Hoque 和 Carenini，2015年），但它们采用了其他方法，例如非负矩阵分解和片段引用，并未提供用户所需的完整细化操作集（Lee 等人，2017年）。

我们的工作可以视为对Smith 等人（2018年）的工作的扩展。与这些研究相比，我们的系统不仅提供了一个完全交互式的 HL-TM 系统，整合了各种细化操作，还采用了更以用户为中心的设计，例如完整的细化历史记录和主题词建议功能，以提升用户体验。

3 提出的系统

本节介绍我们的实现和界面设计。

3.1 优化实现方式

我们的系统使用吉布斯采样作为推断技术，并采用Kumar 等人(2019年) 所述的约束方法来注入新信息。每当用户向系统提供反馈时，系统首先忘记表格-单词分配t_{ji}，然后使用潜在主题的势函数将新信息注入系统 $f(k,w,j)$ (Yang 等人,2015年)，其中 k 是文档j中单词 w 的。方程（1）于是变为：

$$p(t_{ji}=t \mid t^{-ji},k) \propto \mathbb{1}(w_{ji},k_{jt})n_{jt}^{-ji}f_{k_{jt}}^{-w_{ji}}(w_{ji})f(k_{jt},w_{ji},j) \quad (4)$$

先前的研究(Lee等人，2017年；Smith等人，2018年)发现，用户通常更喜欢简单的细化-

操作，因此，我们实现了以下六种在先前研究中用户常用的细化操作：

添加单词 x 到主题 z。我们更新势函数 $f(k,w,j)$，使得当 $k=z$ 和 $w=x$时，$f(k,w,j)$ 为 $=1$；否则将其赋值为 0。

从主题中移除单词 x from topic z. 我们更新势函数 $f(k,\ w,\ j)$，使得当 $f(k,\ w,\ j)=0$if $k=z$ and $w=x$ 时，其值为 1；否则赋值为 1。

交换主题 z 中 w_1 和 w_2 的词序，使 w_2 的顺序高于 w_1。我们首先计算比率 r，即差值n_{w_1}、$z-n_{w_2}$、z与 n_{w_2}之间的比值，其中n_{w_1}、z 和n_{w_2}、z 分别是主题 z中 w_1 和 w_2 的计数，而 n_{w_2} 是除 z之外所有主题中 w_2 的计数。然后我们更新势函数 $f(k,w,j)$，使得当 $k=z$和 $w=w_2$时，$f(k,w,j)$为 $=1$；否则将其赋值为 δ，其中当 $r>1$ 时为 $\delta=0$，否则为$\delta=1.0-r$。

移除文档 d从主题 z中。我们更新势函数 $f(k,\ w,\ j)$，使得$f(k,\ w,\ j)=0$当 $k=z$和 $j=d$时；否则将其赋值为 1。

合并主题 t_2 至 t_1。在下一轮吉布斯采样迭代中，我们只对上一轮吉布斯采样迭代中分配给 t_2 的单词进行采样。我们更新势函数 $f(k,\ w,\ j)$，使得 $f(k,\ w,\ j)=1$当 $k=t_1$和 $w\in t_2$时；否则将其赋值为 0。

拆分主题 t 为两个主题，使用种子词s，即需要将 s 从 t 移至一个新主题t_n。为此，我们首先使用非参数模型创建一个新主题。然后，我们应用添加单词 操作，将 s 中的所有单词添加到 t_n。

3.2 主题词建议

我们在 QD-TM 中开发了一项新颖的主题词建议功能，让用户决定哪些词语应出现在主题词中，哪些不应出现。仅使用主观输入可能会产生虚假的主题词，但结合建议词可以帮助用户引导模型朝正确的方向发展。主题词建议功能集成在模型的吉布斯采样过程中，分为两个阶段。首先，对文档是否与某个主题相关进行采样。其次，根据相关文档更新建议词。

文档相关性采样 对于每个主题k，如果某个文档包含任何建议词，且在上一次吉布斯采样迭代中被采样为与该主题相关，则假定它在当前吉布斯采样迭代中仍与该主题相关。否则，我们使用由 Wang 等人(2016)修改而来的以下公式来判断该文档是否与该主题相关：

$$p(r_k=c\mid \mathbf{r},\pi,\beta,\gamma_{\mathbf{r}},\gamma_{\mathbf{ir}})\propto \begin{cases}(C_c^{R(-m)}+\gamma_r)\times\frac{\prod_v^V\Gamma(n_{k,v}^{-m}+\beta)}{\Gamma(\sum_v^V(n_{k,v}^{-m}+f_{c,m,v})+V\beta)} \\ \text{if } c=1 \\ (C_c^{R(-m)}+\gamma_{ir})\times\frac{\prod_v^V\Gamma(n_{\tilde{k},v}^{-m}+\beta)}{\Gamma(\sum_v^V(n_{\tilde{k},v}^{-m}+f_{c,m,v})+V\beta)} \\ \text{if } c=0\end{cases} \tag{5}$$

其中，$C_c^{R(-m)}$是在相关性状态为 c 的文档数量，不包括当前文档m，$n_{k\ v}^{(-m)}$，是目标主题 k中词项 v 的计数，不包括当前文档中的词语 m，$f_{c,\ m,\ v}$ 是相关性状态为 c 的文档 m 中词项 v 的频率，而 π 表示相关性状态上的伯努利分布。β、γ_r和 γ_{ir}是超参数，而 V 是数据集的词汇量。

自动关键词扩展为了在每次吉布斯采样迭代中获取主题的推荐词语，我们首先计算词项 v 在该主题相关文档中的得分：

$$Score(v)=P_R(v)log\frac{P_R(v)}{P_C(v)}, \tag{6}$$

其中，$P_R(v)$表示相关文档中词项 v 的概率，$P_C(v)$表示整个语料库中词项 v 的概率。我们提取得分较高的词项作为候选词项。然后，计算每个候选词项与主题嵌入之间的余弦相似度，仅当相似度大于0.5时，才将词项添加到建议词列表中。主题 k 的嵌入可通过以下方式获得：

$$emb(k)=\sum_m^M P^i(k,m)emb(m), \tag{7}$$

其中，$p^i(k,\ m)$ 表示第 i^{th} 次迭代中主题 k 的第 m^{th} 个单词的概率；M 表示该主题中代表性词语的总数，通常设为 10；$emb(m)$ 表示该 m^{th} 个单词的预训练词嵌入。

3.3 界面设计

该系统的用户界面由两个窗口组成（附录 A）。在第一个窗口（图A1）中，用户可以定义主题模型的超参数，例如主题的初始数量。如果用户感兴趣的是描述语料库中特定概念或方面的主题，也可以在此窗口中定义主题模型的先验知识。如果未提供先验知识，模型的行为就与不含任何先验知识的传统主题模型相同。

第二个窗口（图 A2）展示训练后模型的详细信息。用户可以通过主题列表查看模型。用户还可以使用系统实现的细化操作来改进模型，并追踪模型变更历史。与以往的系统不同，用户在那些系统中只能按顺序细化模型，无法撤销所做的更改；我们加入了一个采用新颖模型树结构的模型历史面板，让用户可以针对同一个模型节点尝试不同的细化操作，并比较生成的模型。为了帮助用户理解每个主题，我们还展示了自动主题标注算法生成的一组主题标签（Mei 等人，2007）；主题标签是概括主题核心思想的短语。界面的详细说明见附录 A。

4 评估

评估分为两个部分。在第一部分中，我们通过一项受控的实验室实验评估了主题词建议功能的性能，并比较了系统的两个版本。在第二部分中，我们将 HT-LM 系统应用于两个真实场景下的主题建模用例，以评估其性能。

4.1 主题词建议评估

数据集选择了三个常用于主题建模的数据集：20newsgroup[1] 数据集，包含来自 20 个类别的 1.8 万条新闻帖子；TagMyNews [2] 数据集，包含来自 7 个类别的 3.2 万条英语短新闻；以及 SearchSnippets（Xu 等人，2017年）数据集，包含来自 8 个类别的 1.2 万条英语短新闻。

基线 我们的 HL-TM 系统使用 QD-TM 作为底层模型。为了测试主题词语建议 功能能否改进原始模型，我们仅使用此功能对 QD-TM 进行优化，并将优化后的模型与原始 QD-TM 进行比较。

参数化 我们重点关注 QD-TM 的目标主题建模能力。为进行公平比较，我们采用了 Fang 等人 (2021年 中描述的相同实验设置。我们使用查询短语来表示数据集中每个类别的主要概念，并遵循 Fang 等人 (2021年 中相同的设置。我们从数据集中移除了名称无意义的类别，例如 `talk.politics.misc` （位于 20Newsgroup 数据集中）。我们通过 HL-TM 界面的第一个窗口将查询短语整合到模型中。

我们使用 HL-TM 系统对原始 QD-TM 进行了改进，并且仅使用了 添加词语 改进操作。在每次吉布斯采样迭代中，对于每个主题，我们将所有建议词添加到该主题中。对于方程（5）中的超参数， β设置为 0.5，而 γ_r 和 γ_{ir} 均设置为 1。我们进行了 2000 次吉布斯采样迭代。

结果 我们根据主题连贯性和文档检索性能评估了最终模型的质量，评估方法与 Fang 等人 (2021年）一致。Precision@K 得分越高，表明学习得到的主题越具区分性和代表性。一致性得分越高，表明主题的可解释性越好。表 1展示了向模型添加建议主题词后的性能。我们观察到，与原始模型相比，改进后的模型取得了更高的一致性得分。这表明建议词在语义上与各主题相关，有助于生成更加连贯的主题。更高的 Precision@K 得分也表明，建议主题词与数据集的预定义类别高度相关。这是预期结果，因为我们的词语建议功能倾向于识别与相应主题类别相关的更重要词语。我们的系统并不是盲目地将建议词添加到模型中，而是允许用户决定应添加哪些建议词。这样，用户便可以利用其领域知识进一步过滤掉噪声。

4.2 实验室评估

描述 为了将我们的系统与最先进的人机协同主题建模

[1] http://qwone.com/ jason/20Newsgroups/ [2] http://acube.di.unipi.it/tmn-dataset/

模型	20news		TagMyNews		SearchSnippets	
	一致性	Precision@K	一致性	Precision@K	一致性	Precision@K
QD-TM	0.445	0.612	0.413	0.736	0.475	0.741
QD-TM +词语建议	**0.482**	**0.634**	**0.477**	**0.763**	**0.481**	**0.825**

表1：添加建议主题词进行模型优化的平均结果。每个模型运行了五次。我们使用了主题一致性指标 C_V，来自 Röder等人（2015年）

系统（Smith 等人，2018年）进行比较，我们通过大学校园电子邮件列表招募了20名参与者，开展了一项小规模用户研究。其中一组使用与 Smith 等人（2018年）相当的系统——即我们的系统，但不含模型历史、单词建议或主题标注功能（旧系统）；另一组则使用功能完整的新系统（新系统）。在包含两个自变量的因子设计中，我们使用了两个语料库：语料库A由讨论在线教学平台的 Reddit 帖子组成，语料库B由讨论2021年联合国气候变化大会的 Twitter 推文组成。参与者被随机分配到两个群组。第一组使用旧系统处理语料库A，并使用新系统处理语料库B；第二组则使用旧系统处理语料库B，并使用新系统处理语料库A。任务是对这些数据集进行定性分析。

参与者随机从旧系统或新系统开始，以消除训练效应的影响。随后，我们评估了两种系统条件下的平均主题质量。除了比较不同系统之间的主题质量外，我们还询问了参与者对新功能的喜爱程度。所有参与者都能流利使用英语，且不是主题建模领域的专家。每位参与者都会收到一张价值20英镑的亚马逊礼品卡，作为实验薪酬。

数据集与主题模型 我们在实验中使用了两个数据集。Reddit 数据集包含 9,651 篇关于在线教学平台的帖子。这些帖子是从Alqazlan 等人（2021年）提供的原始数据集中随机抽取的。Twitter 数据集包含 8,990 条与 2021 年联合国气候变化大会相关的推文。我们使用与该大会相关的关键词，搜索了 2021 年 10 月 31 日至 2021 年 11 月 17 日期间发布的推文。这些关键词（见表B1）由研究 COP26 气候变化推文的社会科学家提供，并经过数轮人工检查。我们使用 QD-TM 作为底层模型，并采用标准停用词列表和 2000 次吉布斯采样次迭代来初始化主题。根据我们以往处理这些数据集的经验，我们分别将 Reddit 数据集和 Twitter 数据集的初始主题数设为 10 和 13。在任务期间的每次后续更新中，均执行 10 次吉布斯采样迭代。

流程 每次会话平均约持续60分钟，并以面对面的方式进行。我们首先向参与者介绍主题建模，以及如何使用工具优化主题。对于两种系统，我们都要求参与者先阅读每个主题排名前五的帖子，以理解其潜在主题；然后阅读相应的前10个主题词，并使用所提供的优化操作对主题进行优化，直到前10个主题词和前5个帖子与所理解的主题一致。对于新系统，参与者还可以使用新功能——模型历史、词语建议和主题标注，以协助进行细化。我们要求参与者在为某个主题添加所有必要的细化后，点击应用细化 按钮，以更新底层模型。他们可以撤销任何尚未应用到底层模型的操作。完成优化任务后，我们要求参与者评价他们对所得主题的满意程度，以及对新功能的喜爱程度。由于时间限制，我们只要求参与者优化每个系统的前五个主题。每个语料库的初始模型中的前五个主题见表 B2和表B3。

研究结果 我们记录了用户使用工具时的交互。用户使用两个系统共进行了 1,284 次细化。使用频率最高的四种操作包括 删除词语操作（使用了 554 次）、交换词语 操作（使用了 512 次）、添加词语 操作（使用了 171 次），以及删除文档 操作（使用了 81 次）。这与Smith 等人(2018年的研究结果一致。他们也发现，这些操作是使用频率最高的四种操作（不包括添加停用词操作，因为我们的系统未实现该操作）。

对于新系统，添加词语 操作总共执行了 115 次，其中 56 次是添加建议词，这表明建议词语操作的使用频率较高。我们在表B4中报告了各受试者组细化操作的使用情况。

为评估使用新系统是否能够生成比旧系统更优质的主题，参与者需要使用五点量表（1 表示非常不满意，5 表示非常满意）评价他们对两个系统最终主题的满意度。对于 Reddit 数据集，新系统的平均满意度得分为 4.2（SD=0.67），而旧系统的得分为 3.89（SD=0.78）。尽管曼-惠特尼 U 检验（U=30.5，z=0.83887，p=.200）显示这一差异不具有统计显著性，但 10 名参与者中有 8 名对新系统的最终主题感到满意。对于 Twitter 数据集，新系统的得分为 4.0（SD=0.71），而旧系统的得分为 3.2（SD=0.67）。曼-惠特尼 U 检验（U=18.5，z=1.8985，p=.029）显示该差异具有统计显著性，表明新系统有助于提升主题质量。

为了评估模型树功能是否能够帮助用户跟踪变化，我们要求参与者针对以下陈述评价其认同程度："我能够记住更新前模型的样子"，采用五点量表（1 表示强烈不认同，5 表示强烈认同），分别对旧系统和新系统进行评分。旧系统的平均认同度为 3.1（SD=1.07），新系统为 3.9（SD=0.85）。曼-惠特尼 U 检验（U=116，z=2.25868，p=.012）表明，两者之间的差异具有统计学意义。我们还要求参与者针对以下陈述评价其认同程度："新系统的模型树功能能够帮助我跟踪所做的更新"。该陈述的平均认同度为 4.35（SD=1.04）。这有力地表明，新系统的模型树功能能够帮助用户跟踪其所做的更改。为了评估词语建议功能的实用性，我们要求参与者针对以下陈述评价其认同程度："新系统的建议词语功能能够帮助我识别主题中的相关词语"。该陈述的平均认同度为 4.1（SD=0.64）。20 名参与者中有 17 名表示该功能非常实用，这表明该功能是向主题添加词语的一项良好补充。这一发现与观察结果一致。

对新系统中用户交互的观察表明，新增的词语中约有 50%（115 个中的 56 个）来自建议词列表。对于主题标注功能，我们询问参与者是否同意以下陈述："功能完整的系统所提供的主题标签能够帮助我理解主题的含义"。该陈述的平均得分为 3.25 分，20 名参与者中有 11 名指出，部分主题所提供的标签完全不相关。这表明该算法（Mei 等人，2007）与数据集的匹配程度不高，也说明未来需要开发更准确的主题标注算法。

建议 尽管用户研究显示出令人鼓舞的结果，但参与者也提出了一些改进新系统的建议。六位参与者提到，交换词语操作不太符合他们的需求。他们更希望允许将选中的词语插入新的位置，而不是交换两个选中词语的顺序，这样在对主题词进行排序时能够提供更大的灵活性。一位参与者建议，允许直接撤销之前添加的任意操作，而不是只能从最近执行的操作开始撤销。

4.3 案例一：导师在零工经济中的经历

描述 为了在真实的文档分析任务中评估我们的系统，我们邀请了一位正在研究零工经济的研究人员使用我们的模型。在此前的研究中，Alqazlan 等人(2021年) 将潜在狄利克雷分配应用于一个 Reddit 数据集，以识别与零工经济中导师的经历相关的主题。该数据集收录了导师们在在线教学平台上工作经历的相关帖子。我们请这位研究人员使用我们的系统，从数据集中识别相关主题，并评估使用我们的系统能否获得比潜在狄利克雷分配更好的结果。鉴于 Alqazlan 等人 (2021年) 此前发现，包含17个主题的潜在狄利克雷分配模型提取出的相关主题数量最多，为了确保比较的公平性，我们也将系统的初始主题数量设定为17。我们鼓励这位研究人员使用系统提供的任意细化操作，直至她认为每个检测到的主题中排名前5的相关帖子都围绕一个主要主题展开。

结果 在细化过程中，共训练了 16 个模型并创建了两个分支：模型 11、12 和 13 构成一个分支，模型 14、15 和 16 构成另一个

分支。这两个分支均从模型 10 扩展而来，而模型 10 是从初始模型按顺序训练得到的。通过使用模型树面板比较这些模型，研究者发现模型 16 的结果最令人满意。在所有细化操作中，使用次数最多的是移除文档，使用次数最少的是交换词序 。研究者评论说，如果交换词序 仅调整主题中某个单词的位置，而不是将一个单词与另一个单词互换，那么它可能会更有用。

我们将潜在狄利克雷分配与模型 16 进行了比较，并在此呈现定性结果。对于这两个模型，研究人员都需要为每个主题手动分配标签。潜在狄利克雷分配模型中，有 11 个主题被认为与导师在零工经济中的经历相关，其中两个（表 C1）未被模型 16 识别；而在模型 16 中，有 16 个主题被认为相关，其中七个主题（表C2）未被潜在狄利克雷分配识别。这表明，我们的 HL-TM 系统能够帮助研究人员识别出更多相关主题。

为确定使用我们的系统能否帮助生成质量更高的主题，我们测量了潜在狄利克雷分配和模型 16 共同识别出的 9 个相关主题的主题一致性得分（表 C3）。我们采用了基于外部语料库（维基百科）的表现最佳的主题连贯性测量指标 CV（Röder等人，2015年）。我们发现，我们的系统在六个主题上得出的主题连贯性得分更高，而潜在狄利克雷分配仅在三个主题上得分更高。这一发现与研究人员的看法一致：潜在狄利克雷分配得出的主题中，只有两个的质量优于模型 16。附录 C 列出了潜在狄利克雷分配和模型 16 的相关主题。

4.4 用例二：针对预先确定的专利价值类别制定专利策略

描述 我们邀请了第二位研究人员对我们的系统进行评估，该研究人员从事专利策略研究。在此前的一项研究中，Ribeiro和Shapira（2020年）确定了一组用于合成生物学专利的专利价值类别（表D1）。然而，他们的工作完全基于人为评估，因此耗时较长，并且可能出现有偏差的评估，所以只分析了 102 份专利文档。在这一使用案例中，该研究人员旨在使用我们的系统，将更大的专利样本归入预定义类别，并减少人为选择偏差。2607 份相关专利来自美国专利商标局（美国专利商标局）数据集。预定义类别见附录 D。

为了识别数据集中属于预定义类别的文档，研究者使用界面的第一个窗口融入先验知识。附录 D 的表 D2 列出了各类别的最终概念词。初始主题数设为 20。

结果 通过使用用户界面的第一个窗口融入先验知识，初始模型能够粗略推断出预定义类别对应的主题，尽管这些主题的质量尚不高。随后，我们请研究者查看这些主题排名前 10 的相关文档，以评估分类质量。平均精确率为 0.3。之后，研究者使用该系统细化这些主题，直到她认为系统能够将排名前 5 的相关文档准确归入各类别，并且对生成的主题词感到满意。总共训练了 23 个模型，并创建了五个分支，其中模型23是最佳模型。细化之后，我们请研究者进一步查看细化后主题排名前 10 的相关文档，并与初始模型进行比较。细化后主题的平均精确率为 0.96，远高于初始模型的结果。这证明，与以往的人工评估方法相比，使用该系统有助于识别出更多属于预定义类别的文档。在总计 62 次操作中，添加单词操作被执行了 62 次，其中 41 次涉及添加系统建议的词语。这表明，我们的主题词建议功能确实能够识别与特定主题高度相关的词语。

我们还比较了初始模型中的聚焦主题（表 D2）和模型23中的聚焦主题（表 D3）的主题一致性得分。得分分别为 0.404 和 0.434。这表明，使用该系统有助于生成质量更高的主题。

5 结论与未来工作

我们开发了一种新颖的、以用户为中心的交互式 HL-TM 系统，以解决先前工作的局限性（Smith 等人，2018年）。系统提供了先进的用户界面和模型历史面板，方便用户比较不同模型并撤销不恰当的更改。采用 QD-TM 作为底层模型，使系统同时具备全分析和定向分析能力。此外，

系统还集成了一种新颖的主题词建议功能，评估结果表明，该功能有望为主题推荐合理且连贯的词语。一项小规模实验性用户研究和两个详细的用例进一步验证了该系统在现实任务中的实用性。从这些用例中，我们观察到，两位研究人员都训练了许多不同的模型，并创建了多个分支（用例一训练了 16 个模型并创建了两个分支；用例二训练了 23 个模型并创建了五个分支）。这表明，研究人员在优化主题模型时很可能改变想法，因此，让研究人员比较不同模型并撤销更改很有帮助。未来，我们计划根据用户评估结果改进系统，并开展更广泛的评估，以收集有关在各种应用场景中部署该系统的反馈。

局限性

在用户评估中，我们使用了两个小型数据集，因此参与者不会受到延迟问题的影响，即用户无需在算法执行更新时等待。由于 QD-TM 比潜在狄利克雷分配更加复杂，而且单词建议功能进一步增加了计算的复杂性，因此大型数据集可能会导致延迟问题。我们系统的时间复杂度为 O(DLK+ KV)，其中 D 表示所使用的文档数量，L 表示文档的平均长度，K 表示主题数量，V 表示词汇量。我们使用两种不同规模的数据集（20,000 条推文与 8,990 条推文）对系统进行了测试。使用 20,000 条推文时，用户交互过程中模型更新的平均等待时间（10 次吉布斯采样迭代）为 34 秒，而使用 8,990 条推文时为 14 秒。

根据 Smith 等人 (2018年)，较长的等待时间会对用户使用交互式系统产生负面影响，这表明随着数据集规模增大，延迟可能会成为我们系统的一个问题。然而，众所周知，提供时间上的宽裕度可以减轻这些影响（Conn，1995年）。我们将在未来解决这一问题。

除了延迟问题外，我们开展的实验室评估也存在局限性。并非所有参与者都熟悉定性分析。很可能，那些拥有丰富定性分析经验的参与者能够发现更有价值的模型细化，从而得出质量更高的主题。

致谢

这项工作部分得到英国工程与物理科学研究委员会资助（资助编号：EP/V048597/1）。YH 获得由英国研究与创新署资助的图灵人工智能研究员奖助金（EP/V020579/1）。ZF 获得由华威大学和中国国家留学基金管理委员会共同资助的博士生奖学金。

参考文献

Lama Alqazlan、Rob Procter 和 Michael Castelle。2021年。自然语言处理与数字人文研讨会，第*18*届国际自然语言处理会议。

David M Blei、Andrew Y Ng 和 Michael I Jordan。2003年。潜在狄利克雷分配。《机器学习研究杂志》，3：993–1022。

Jordan Boyd-Graber、David Mimno 和 David Newman。2014年。主题模型的维护与培育：问题、诊断与改进。*Handbookofmixed membershipmodelsandtheir applications*，225255。

Ryan C Boyer、William T Scherer、Cody H Fleming、Casey D Connors 和 N Peter Whitehead。2017年。一种用于研究复杂语料库中系统思维的人机协同方法。*IEEESystemsJournal*，12(3):2937–2948。

Jonathan Chang、Sean Gerrish、Chong Wang、Jordan L Boyd-Graber 和 David M Blei。2009年。《读懂茶叶：人类如何解读主题模型》。载于*Advances in Neural Information Processing Systems*，第288–296页。

Jaegul Choo、Changhyun Lee、Chandan K Reddy 和 Haesun Park。2013年。UTOPIAN：基于交互式非负矩阵分解的用户驱动的主题建模。*IEEE Transactions on Visualization and Computer Graphics*，19(12)：1992–2001。

Alex Paul Conn. 1995年. 时间可供性：诊断性可用性启发式准则中的时间因素。载于*SIGCHI*人机交互系统中的人为因素会议论文集，第186–193页。

Stephan A Curiskis、Barry Drake、Thomas R Osborn 和 Paul J Kennedy。2020年。对两个在线社交网络 Twitter 和 Reddit 中的文档聚类与主题建模进行评估。*InformationProcessing & Management*，57(2):102034。

Ramakrishna Dantu、IndikaDissanayake 和 Sridhar Nerur。2021年。物联网的探索性分析

在医疗保健中的应用：一种主题建模与共引方法。信息系统管理，38(1):62–78。

郑方、Yulan He 和 Rob Procter。2019年。一种查询驱动主题模型。载于计算语言学协会论文集：*ACL-IJCNLP 2021*年会议论文集，第1764–1777页，线上。计算语言学协会。

Barney G Glaser、Anselm L Strauss 和 Elizabeth Strutzel。1968。扎根理论的发现：定性研究策略。护理研究，17(4):364。

Nihit Goyal 和 Michael Howlett。2021年。“衡量”新冠病毒-19政策应对措施的组合：使用主题建模进行比较政策分析。比较政策分析杂志：研究与实践，23(2):250–261。

Tobias Heidenreich、Fabienne Lind、Jakob-Moritz Eberl 和 Hajo G Boomgaarden。2019年。“欧洲难民危机”的媒体框架动态：一种比较主题建模方法。难民研究杂志，32(特刊_期_1):i172–i182。

Gregor Heinrich。2009年。主题模型的一种通用方法。载于机器学习与数据库知识发现联合欧洲会议，第517–532页。Springer。Enamul Hoque 和 Giuseppe Carenini。2015年。Con- visit：用于探索异步线上对话的交互式主题建模。载于第*20*届智能用户界面国际会议论文集，第169–180页。Jagadeesh Jagarlamudi、Hal Daumé III 和 Raghaven- dra Udupa。2012年。将词汇先验纳入主题模型。载于欧洲计算语言学协会第*13*届会议论文集，第204–213页。Varun Kumar、Alison Smith-Renner、Leah Findlater、Kevin Seppi 和 Jordan Boyd-Graber。2019年。你为什么不听我的？比较人类参与主题模型中的用户控制。*arXiv* 预印本 *arXiv:1905.09864*。Tak Yeon Lee、Alison Smith、Kevin Seppi、Niklas Elmqvist、Jordan Boyd-Graber 和 Leah Findlater。2017年。人性化的一面：非专家用户如何感知、理解并修正主题模型。国际人机交互研究期刊，105：28–42。Jeffrey Lund、Connor Cook、Kevin Seppi 和 Jordan Boyd-Graber。2017年。串联锚定：一种用于交互式主题建模的多单词锚定方法。载于计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），第896–905页。

Qiaozhu Mei、Xuehua Shen 和 ChengXiang Zhai。2007。多项式主题模型的自动标注。载于第*13*届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集，第490–499页。

David Mimno、Hanna Wallach、Edmund Talley、Miriam Leenders 和 Andrew McCallum。2011 年。优化主题模型中的语义一致性。载于*2011* 年自然语言处理实证方法会议论文集，第 262–272 页。

Barbara Ribeiro 和 Philip Shapira。2020年。创新的私人价值与公共价值：合成生物学的专利分析。研究政策，49(1):103875。

Michael Röder、Andreas Both 和 Alexander Hinneb- urg。2015年。探索主题连贯性度量的空间。载于 第八届 *ACM* 国际网络搜索与数据挖掘会议论文集，第 399–408 页。

Carson Sievert 和 Kenneth Shirley。2014年。Ldavis：一种可视化和解释主题的方法。载于交互式语言学习、可视化与界面研讨会论文集，第63–70页。

Alison Smith、Varun Kumar、Jordan Boyd-Graber、Kevin Seppi 和 Leah Findlater。2018年。闭环：用户中心设计与人机协同主题建模系统的评估。载于第*23*届智能用户界面国际会议，第293–304页。Yee Whye Teh、Michael I Jordan、Matthew J Beal 和 David M Blei。2006年。层次狄利克雷过程。美国统计学会期刊，101(476)：1566–1581。Di Wang、Marcus Thint 和 Ahmad Al-Rubaie。2012年。半监督潜在狄利克雷分配及其在文档分类中的应用。载于*2012*年 *IEEE/WIC/ACM* 网络智能与智能体技术国际会议，第3卷，第306–310页。IEEE。Shuai Wang、Zhiyuan Chen、Geli Fei、Bing Liu 和 Sherry Emery。2016年。面向目标的主题建模：聚焦分析。载于第*22*届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，第1235–1244页。Jiaming Xu、Bo Xu、Peng Wang、Suncong Zheng、Guanhua Tian 和 Jun Zhao。2017年。用于短文本聚类的自训练卷积神经网络。神经网络，88：22–31。Yi Yang、Doug Downey 和 Jordan Boyd-Graber。2015年。将知识融入主题模型的高效方法。载于*2015*年自然语言处理实证方法会议论文集，第308–317页。

附录 A：界面设计

该系统的用户界面由两个窗口组成。如果用户对描述语料库中特定概念或方面的主题感兴趣，可以在第一个窗口中定义主题模型的先验知识（图A1）。他们可以使用查询短语来定义感兴趣的概念。系统会使用Fang 等人 (2021年）提出的概念词提取器，将输入的查询短语扩展为一组候选概念词。用户可以运用自己的知识，决定哪些词语应纳入主题，并点击旁边的“+”按钮，将其添加到该主题的最终概念词列表中。用户也可以根据自己的知识，将候选词列表中没有的词语添加到最终列表中。点击相应的“-”按钮即可从最终列表中移除词语。用户还可以将查看选项从“按关键词”切换为“按文档”，以查看与输入查询相关的文档。

如果用户对语料库中的主题没有特别偏好的兴趣，可以将每个主题的最终概念词列表留空，此时模型会像没有任何先验知识的传统主题模型一样运行。用户可以点击“应用”按钮，将设置应用于底层模型。用户可以点击底部的“查看”按钮查看生成的主题，该按钮会将用户带到用户界面的第二个窗口。

界面的第二个窗口（图 A2）包含三个面板：模型历史面板（左侧面板）、模型详情面板（中间面板）和主题间距离图（右侧面板）。模型历史面板由两个子面板组成：用于以树形结构显示细化后模型树的面板（上方面板），以及细化历史面板（下方面板）。每当用户细化一个选定模型时，树中都会新增一个模型。如图A2所示，用户还可以从同一个模型节点创建新的分支。例如，模型 5、模型 6 和模型 7 都是从模型 4 细化而来的。细化历史面板允许用户点击两个相连模型之间的边，查看它们之间的细化历史。用户还可以查看选定模型尚待处理的细化，如图A2所示。

在用户点击“应用细化”按钮之前，待处理的细化不会应用到底层模型。我们还提供了一个“撤销”按钮，以便用户撤销先前添加的细化。模型历史 面板允许用户进行操作。比较此前细化过的模型，并跟踪其先前的更改，以进一步辅助细化过程。点击“下载模型”按钮，用户可以将整个模型树下载到本地计算机；点击“加载模型”按钮，则可以加载先前下载的模型树并继续细化。

该模型详情面板位于图A2的中部，概述了选定模型的主题以及每个主题中的高频词。权重表示该主题在整个语料库中的普遍程度。用户可以在“主题”列中输入新名称来重命名主题。用户还可以在此面板中合并或拆分主题。图A2 的右侧是主题间距离图。每个圆圈代表一个主题，其大小表示该主题在语料库中的权重。主题间距离图是直观揭示主题模型质量的一种方式，主题之间的距离越大，表示模型越好（Sievert 和 Shirley，2014年）。点击窗口中的“查看”按钮后，选定主题的具体信息将如图A3所示呈现。

图A3 左上方显示了自动主题标注算法生成的一组主题标签（Mei 等人，2007年），其中主题标签是概括主题主要思想的短语。主题标签旨在帮助用户理解主题。我们还展示了该主题的高频词及相应的主题词权重。权重表示词语在主题中的普遍程度。用户可以在此处应用添加、删除或交换词序等细化操作。图A3右侧显示了与该主题相关的高相关文档，并根据其权重对文档进行排序，该权重表示文档对主题的普遍程度。用户可以在此处应用移除文档细化操作。点击“查看”按钮后，用户可以查看选定文档的详细信息。

当用户点击图 A3中的“添加”按钮时，会出现“添加词语”子面板（图 A4）。面板中会显示由主题词建议功能推荐的词语列表。用户可以从建议词中选择，或根据自身知识添加词语到选定主题。

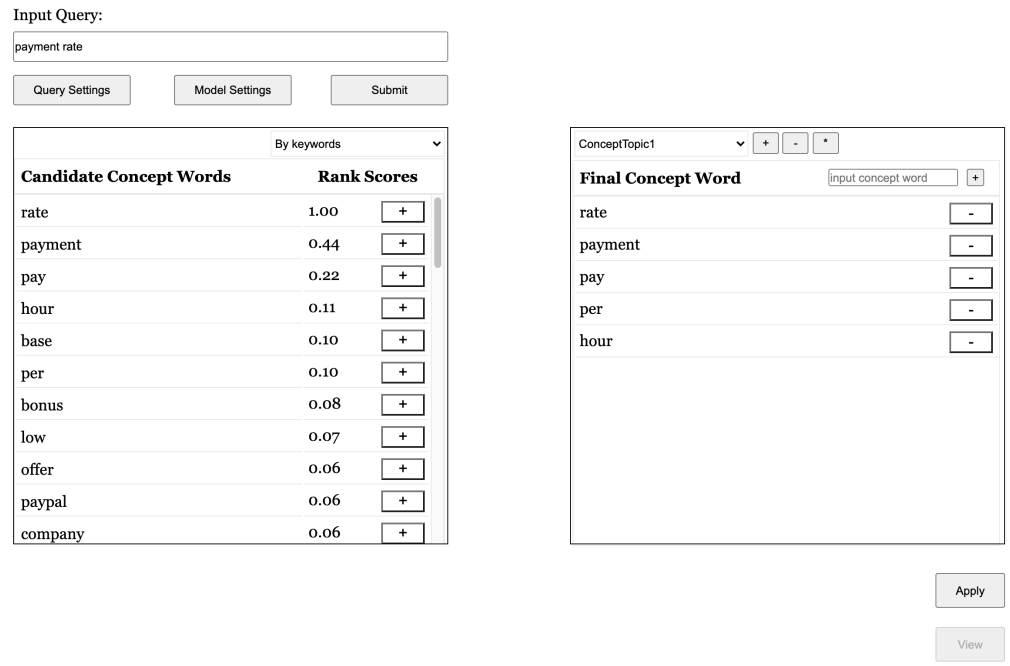

图 A1：用户界面的第一个窗口。用户可以在此定义主题模型的先验知识。

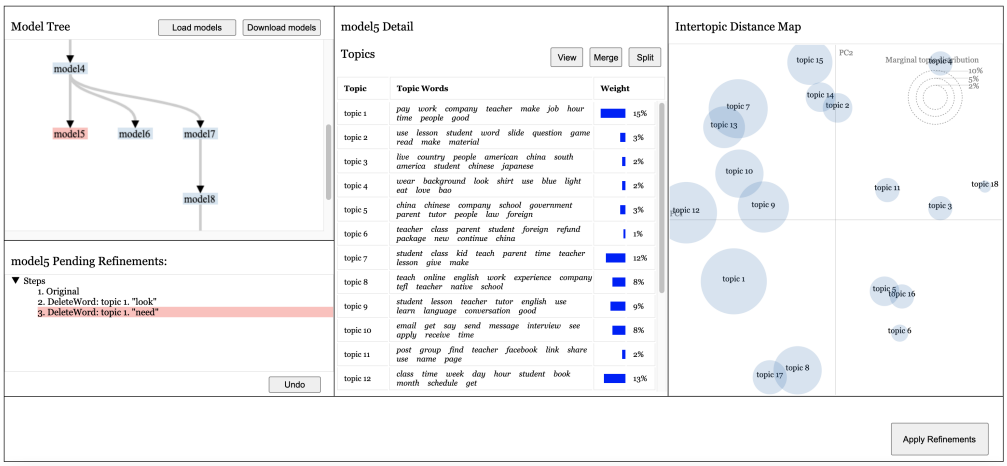

图 A2：用户界面的第二个窗口，用户可以在此查看选定模型的详细信息，并申请合并或拆分主题的细化操作。用户还可以比较此前已细化的不同模型，并跟踪之前所做的更改。

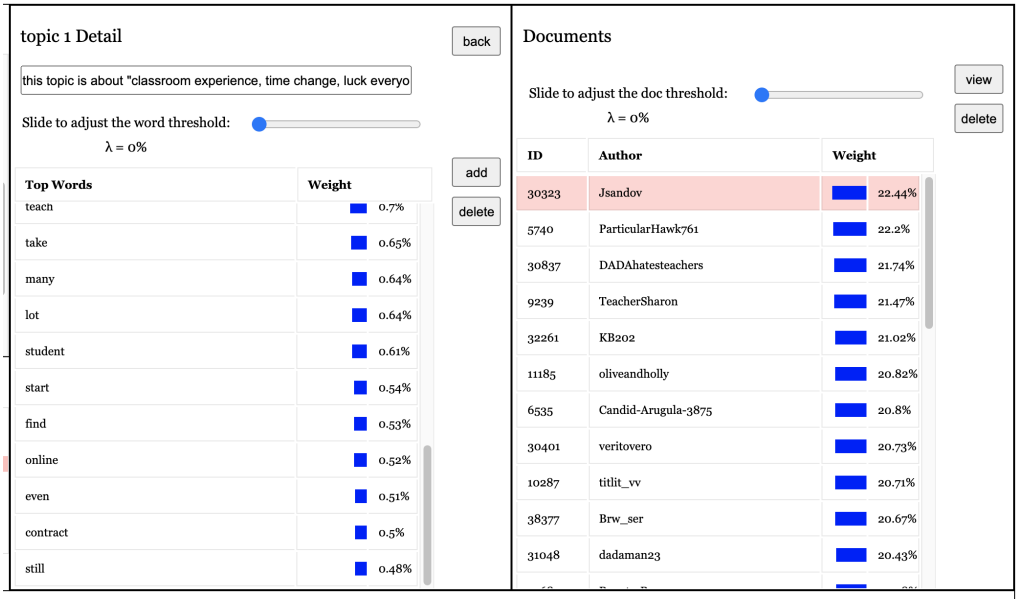

图 A3：选定主题的具体信息。用户可在此申请添加、移除或交换词序等细化操作。

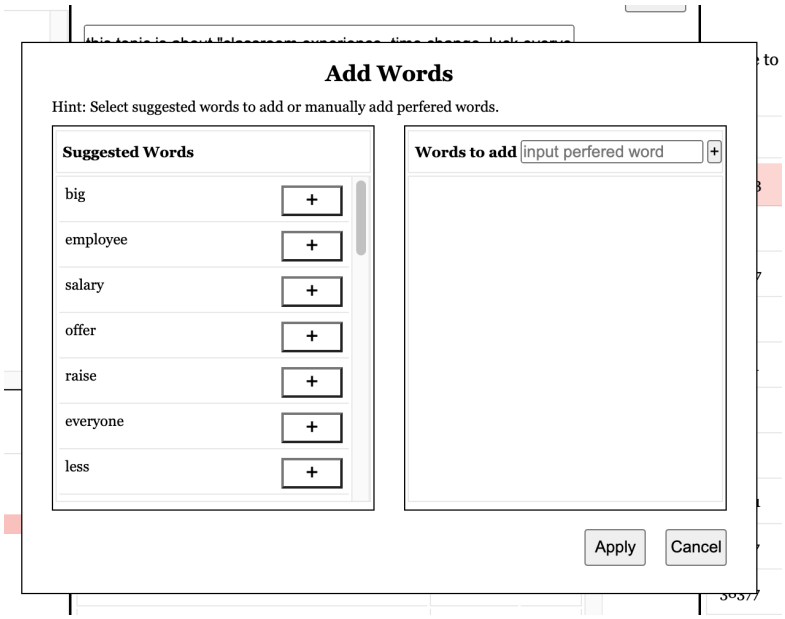

图 A4：“添加词语”子面板。

附录 B：小规模用户研究（第 4.2 节）的表格

关键词
警务联络、警务联络官、PLO、蓝色围兜、#和平抗议、#right-toprotest、警察、警务工作、无政府状态、无政府主义、暴力、#气候正义，#气候危机、#气候行动、cop26、#全球行动日、#GlobalDayfor-ClimateJustice、#漂绿、猪、法西斯分子、冲锋队员、纳粹分子、人群、游行，集会，暴民，极端分子，班级，特权，环保狂热分子，破坏者，疯子，疯子，环保法西斯分子，伪君子，极左派，混乱，噱头，马克思主义者，恐怖，乌合之众，无政府主义者，污秽，警察，小丑，警察，警察，警察，混蛋，觉醒派，环保斗士，运动，激进分子，极端分子，狂热分子，威权主义者，暴政，暴君，笨蛋，笨蛋，人渣，恶心的，糟糕的，难以置信的，邪恶的，令人恐惧的，自私的，恶毒的，暴力的，暴徒，残暴的，恶毒的，打斗，大屠杀，血，伤害，敌意，攻击性，武力，袭击、入侵、冒犯、摩擦、压力、紧张、损害、伤害、危害、阻碍，水壶、包含、逮捕、监禁、收费、谢谢、感谢、狂欢节、节日、乐趣，享受、政党、紧张、紧张的、扰乱、柔软

表 B1：用于检索与2021年联合国气候变化大会相关推文的关键词。

主题	排名前 10 的词语
主题 1	小时 班级 周 薪酬 月 工作 时间 天 开始 制作
主题 2	电子邮件 申请 工作 公司 良好 检查 发送 面试 聘用 流程
主题 3	教学 英语 线上 母语者 英语作为外语教学 学位 经验 公司 教师资格证书
主题 4	学生 班级 给出 评分 课程 反馈 时间 家长 水平 不好
主题 5	导师账户学生课程视频群组个人资料分享银行 PayPal

表 B2：Reddit 数据集初始模型的前五个主题。新旧系统都使用相同的初始模型。

主题	排名前10的词语
主题 1	气候行动 变化 COP 人民 格拉斯哥 世界 今天 天 AMP
主题 2	政党 投票 保守党 工党 约翰逊 鲍里斯 政府 议员 帕特森 丑闻
主题 3	放大排放 国家 气候 碳 能源 天然气 全球煤炭 燃料
主题 4	收费 薪酬 人们 新冠病毒 资金 案例 工作 索赔 给予 电子邮件
主题 5	游戏 很棒 玩耍 团队 今天 享受 放大 乐趣 获胜 时间

表 B3：Twitter数据集初始模型的前五个主题。新旧系统使用了相同的初始模型。

操作	新系统/Reddit	新系统/Twitter	旧系统/Reddit	旧系统/Twitter	总计
删除词语	74	205	198	77	554
重新排列词语	118	152	139	103	512
添加自定义词语	14	45	32	24	115
添加建议词语	7	49	/	/	56
删除文档	12	37	20	12	81
拆分主题	0	3	0	4	7
合并主题	0	1	0	1	2

表 B4：每个受试者组对细化功能的使用情况。“新系统/Reddit”表示使用新系统处理 Reddit 数据集。

附录 C：用例一的表格（第 4.3 节）

标签	排名前 10 的词语
加入或离开某个平台的原因-form	公司、教师、工作、工作、薪酬、人们、制造、资金、线上、思考
与某个平台的沟通不畅-表单管理	知道，任何人，视频，导师，帮助，询问，让，请，申请，面试

表 C1：潜在狄利克雷分配识别出的相关主题，但模型 16 未能识别。

标签	排名前 10 的词语
最低工资和生活工资	薪酬、工作、公司、教师、制作、小时时间、费率、工资、工作、时间
学生的国籍	中国，人民，国家，生活，中国人，学生，美国人，美国，世界，政府
对导师施加的限制可以在辅导时穿着	穿着、背景、衬衫、蓝色、吃、光、使用、包、爱、恐龙
中国新法关于校外辅导	教师、中国、家长、中国人、公司、外国的、学校、班级、新、学生
专业人士之间的差异-导师与社区导师	学生，课程，教师，导师，使用，英语，学习，语言，进行，对话
关于 Facebook 的讨论由平台创建的群组管理	发布，群组，查找，Facebook，链接，查看，教师，使用，分享，名称
导师表达愤怒和不满对平台的满意度	人们，制作，好，操，帖子，屎，去，感觉，娜，对不起

表 C2：模型 16 识别出但潜在狄利克雷分配遗漏的相关主题。

标签	前10个单词（潜在狄利克雷分配）	前 10 个单词（模型 16）
预约和工作时间	周、小时、天、时间、书籍、班级、时段、日程安-排、开放、月	班级、周、时间、天、小时、书籍、学生，日程安排、月、时段
评分系统	学生、教师、工作、班级、给出、评分、重新-盟友，思考，月，公司	学生，评分，通话，分钟，预约，时间，小时, ph, 导师, 周
教学材料和方法-ods	学生、使用、课程、问题、单词、询问、说、制作，学习，对话	单词，使用，游戏，阅读，学生，幻灯片，玩耍，问题，课程，句子
技术问题	班级，分钟，学生，打电话，时间，展示，开始，迟到，发生，预约	使用、问题、工作、应用程序、互联网、麻烦、尝试，手机、电脑、测试
付款	课程、薪酬、时间、学生、小时、教师、费率，导师, base, 分钟	班级、小时时薪、每节课奖金、教学、分钟，base, 课程 ，学生
在班级中与孩子相处的经历	孩子、教学、水平、良好、年长、年、思考、确实，感受、说	学生、班级、孩子、教学、家长、时间、课程、给予、教师、制作
招聘流程	电子邮件、发送、消息、说、获取、群组、回复、返回、尝试、查看	面试、申请、视频、演示、招聘、良好、ap-申请，新公司，流程
职位要求	英语，母语，居住，国家，使用者，语言—语言，中国，工作，美国人，非	教授，线上，英语，母语，英语作为外语教学，公司，工作，经验，教师，学校
银行转账和交易-费用	评分, 账户, 薪酬, 演示, 使用, 银行, 测试, PayPal, 付款, 互联网	账户, 银行, 薪酬, PayPal, 使用, 付款, 转账, Payoneer, 费用, 资金

表C3：潜在狄利克雷分配和模型 16 均识别出的相关主题

附录 D：用例二（第4.4节）的表格

类别	定义
市场和工业机遇-tunities	发明进入现有市场的潜力
成本与效率	通过提高工艺效率降低生产成本
提高化合物产量	基于新型工艺提升化合物生产力
扩大生产规模	将生产提升至商业化水平
科学进展	对知识生产的贡献
环境可持续性	对环境质量与保护的贡献
人类健康	人类健康质量的改善
粮食安全	避免与人类食物来源竞争
动物健康	动物健康质量的改善

表D1：来自（Ribeiro和Shapira，2020年）的类别表。

类别	最终概念词
市场与工业机遇	市场、商业、价值、存在
成本与效率	通过提高工艺效率降低生产成本
提高化合物产量	化合物，产量，增加，植物
扩大生产规模	生产，改进，规模，大
科学进展	进展 ，益处 ，归档
环境可持续性	可再生、可持续的能源
人类健康	健康、人类、患者、癌症、疾病
粮食安全	食物、粮食安全、供应、制备、链
动物健康	动物 ，健康，疾病

表D2：每个预定义类别的最终概念词。

类别	主题（前10个词）
市场与工业机遇	健康的、改进的、血清、水平、变体、商业的、价值、浓度、女性、样本
成本与效率	效率，提高，成本，增加，高效，增强，反应，扩增，低，目标
提高化合物产量	植物，耐受性，改良，化合物，产量，标记，增加，除草剂，大豆，抗性-性
扩大生产规模	改进，改进后的，生产，活性，抗体，酶，特性，多肽，稳定性，表达
科学进展	益处，折叠，有效，更大，组成，领域，尺寸，皮肤，提供，材料
环境可持续性	可再生的，能源，产品，来源，柴油，生产的，碳，脂质，可持续的，燃料
人类健康	疾病、患者、有效、癌症、数量、抗体、人类、治疗、给药、受试者
粮食安全	链、医疗保健、ge、mm、食物、制备、光、健康的、列、地区
动物健康	健康，动物 ，疾病 ，有效的，国家的，研究所，dose，确定的 ，行政管理 ，数量

表D3：由初始模型识别出的聚焦主题 型。

类别	主题（前 10 个词语）
市场与工业机遇	价值、商业、市场、工业、存在、业务、全球、增加、铁调素、年份
成本与效率	效率、改进、增加、增加、成本、增强、反应、高效、扩增、低
提高化合物产量	植物，效率，化合物，产量，增加，转化，改善，基因，抗性，增加
扩大生产规模	提高、生产、改善、改进、提升、更高、活性、更大、表达、酶
科学进展	益处、提供、领域、改善、皮肤、清洁、酶、表面、纤维素酶、材料
环境可持续性	可再生能源、减少、生物燃料、来源、柴油、生物燃料、碳、脂质、产品
人类健康	疾病，患者，癌症，诊断，有效，人类，肝脏，肺，临床，抗体
粮食安全	链、样本、食物、制备、测定、光、供应、改进、供体、培养物
动物健康	健康，动物，医学，疾病，护理，健康，营养，种群，临床，贝塞斯达

表D4：模型23识别出的聚焦主题。