BTVis: 基于 BERTopic 的交互式层次主题建模可视分析系统

卢文昊, 邹杭纳, 汤颖*

(浙江工业大学计算机科学与技术学院 杭州 310023)
(ytang@zjut.edu.cn)

摘 要: 主题建模是自然语言处理中的重要文本挖掘方法, 但其建模过程复杂且会生成部分不符合用户期望的结果. 为帮助非专家用户理解模型建模过程, 高效掌握和修改模型结果, 设计了基于 BERTopic 的交互式可视分析系统. 系统通过以下关键功能提升 BERTopic 的可解释性与实用性: (1) 挖掘并展示 BERTopic 层次聚类的中间过程, 直观地揭示主题生成机制; (2) 分析离群文档, 揭示其与主题间的潜在关系; (3) 提出多粒度局部模型编辑算法, 增强 BERTopic 主题模型准确性; (4) 开发基于 Web 的交互式层次主题模型系统 BTVis, 支持用户通过可视分析与交互探索理解并提升模型结果. 在 TED、豆瓣影评等真实长文本和短文本数据集中进行了定性分析, 面向 100 位参与者设计用户实验, 并通过一致性、多样性和稳定性指标和其他模型进行对比, 实验结果验证了所提系统的有效性和实用性.

关键词: 文本数据; 主题建模; 模型提升; 可视分析
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024-00082

BTVis: An Interactive Hierarchical Topic Modelling Visual Analysis System Based on BERTopic

Lu Wenhao, Zou Hangna, and Tang Ying*

(*College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou* 310023)

Abstract: Topic modelling is an important text mining method in natural language processing, but its modeling process is complex and can generate results that do not always align with user expectations. To address this issue and enable non-expert users to understand the modelling process and modify model results quickly, we propose the interactive visual analysis system BTVis based on BERTopic. The system enhances the interpretability and usability of BERTopic through the following key features: (1) explored and visualized the intermediate process of hierarchical clustering of BERTopic to intuitively reveal the topic generation mechanism; (2) analyzed the outlier documents to uncover potential relationships between the outlier documents and the topics; (3) proposed the multi-granular local model editing algorithms to enhance the accuracy of the BERTopic model; (4) implemented the interactive hierarchical topic modeling system BTVis in a web environment, allowing users to enhance BERTopic results through visual analysis and interactive exploration. We conducted experiments including qualitative analysis, user experiments and quantitative tests with multiple real datasets. The results show the effectiveness and practicality of the system.

Key words: text data; topic modelling; model enhancement; visual analysis

---

收稿日期: 2024-02-09; 修回日期: 2024-12-19. **基金项目**: 浙江省自然科学基金重点项目(LZ23F020010); 国家自然科学基金重大项目(72192820); 国家自然科学基金面上项目(61972355). **卢文昊**(1999—), 男, 硕士研究生, 主要研究方向为数据分析可视化; **邹杭纳**(2002—), 女, 硕士研究生, 主要研究方向为数据分析可视化; **汤颖**(1977—), 女, 博士, 教授, 博士生导师, CCF 会员, 论文通信作者, 主要研究方向为大数据分析与可视化、计算机图形学.

现实世界中的大规模文本信息多种多样, 涵盖人文学科的研究成果、社交媒体内容以及在线评论等, 由于这些信息复杂程度高、规模量大, 因此用户往往难以快速地获取重要信息. 主题建模作为一种高效的无监督学习方法, 可以自动化地分析大规模文本数据, 通过计算主题对内容进行总结概览, 让用户快速、有效地掌握文本数据的主要内容, 已被广泛地应用在自然语言处理(natural language processing, NLP)的各类任务中, 如文本分类、情绪分析和信息检索等.

一般的主题建模模型通过基于词袋的统计模型得到主题分布, 如潜在狄利克雷分配(latent Dirichlet allocation, LDA)[1]和概率潜在语义分析(probabilistic latent semantic analysis, PLSA)[2]等. 随着深度学习技术的发展, 人们将传统主题建模与深度神经网络技术相结合, 提出深度主题建模方法[3]. 与传统模型相比, 深度主题建模方法能够灵活地学习概率向量或词嵌入来表示文档-主题分布和主题-单词分布, 取得的模型质量更高, 泛化能力更强. BERTopic 是一种基于 BERT 模型[4]开发并被广泛使用的深度神经网络主题模型[5], 它首先使用预训练模型计算文档语义嵌入, 然后基于 HDBSCAN 技术[6]进行聚类, 最后通过基于类的 c-TF-IDF 变体提取每个主题簇的关键词, 解决了基于密度聚类和基于中心采样之间的问题, 使 BERTopic 的结果具有更好的解释性; 此外, 该模型适用性强, 在长文本和短文本上都有很好的主题提取能力.

尽管 BERTopic 能够取得很好的主题建模结果, 但仍有以下挑战:

(1) BERTopic 是黑盒模型, 用户无法直观地理解主题建模的中间过程. 在建模过程中, BERTopic 使用 HDBSCAN 技术对文档进行层次聚类, 而聚类过程将直接影响最终生成的主题模型结果. 但是, 目前 BERTopic 仅提供了对模型结果的可视化展示, 不支持用户查看模型聚类的过程, 更不能进行交互式的探索和分析.

(2) BERTopic 在聚类过程中可能会出现错误. 在 HDBSCAN 技术的层次聚类过程中, 每个层次都会产生离群文档. 然而, 受到语义嵌入方法和参数设置的影响, 部分有价值的文档会被错误地判断为离群文档, 这些离群文档被 BERTopic 统一分配至垃圾主题中, 而用户往往难以从垃圾主题中提取有效的信息. 因此, 为用户对离群文档进行挖掘分析提供指导, 优化 BERTopic 的建模结果是有必要的.

(3) BERTopic 不支持局部修改. 目前, BERTopic 只提供简单的主题级别的合并方法, 并没有提供其他级别和更加丰富的修改方式, 如果用户发现了主题内存在的错误并想要优化主题, 只能通过调参更改模型结果. 例如, 当用户认为主题 A 中的一部分文档更应该被分配至主题 B 时, 很难通过调参来实现这一细节的精确改变.

针对上述问题, 本文提出基于 BERTopic 的可视分析系统——BTVis. 首先挖掘 BERTopic 层次聚类的中间过程, 以可视化的形式展示 BERTopic 的层次聚类结果, 通过交互探索的方式, 帮助用户分析主题的生成过程, 增强 BERTopic 的可解释性; 然后以层次结构对离群文档进行分析, 提供可视化交互, 揭示离群文档与主题间的潜在关系, 并根据 BERTopic 的聚类过程, 依照层次结构对离群文档进行更细粒度的主题建模, 为用户重新分配有价值的离群文档提供可视化指导; 考虑 BERTopic 的算法并不适用于修改模型局部结果, 在 BERTopic 的基础上提出了增强 BERTopic 主题模型准确性的多粒度局部模型编辑算法, 设计主题级、文档级和单词级的局部编辑功能多粒度地优化模型结果, 修复错误的文档分类, 挖掘更深入的主题信息, 保证在局部范围内修改错误的主题, 不影响其他主题; 此外, 结合 ChatGPT 自动生成每个主题的标题, 帮助用户快速地浏览主题语义, 筛选感兴趣的主题并找出存在明显错误的主题. BTVis 允许用户从上而下地探索主题结构, 结合主题、文档和单词 3 种层次的辅助视图快速地分析建模结果的准确性, 找出潜在的问题和用户关注的信息, 使用户在使用上更便捷, 更理解和更信任模型的结果. 最后, 实现 Web 环境下的交互式层次主题模型系统 BTVis.

1　相关工作

1.1　主题建模

主题建模可以分为基于词袋的模型和基于神经网络的模型 2 大类.

基于词袋的模型包括概率模型(如 LDA 方法[1])和非概率模型(如非负矩阵分解(nonnegative matrix factorization, NMF)方法[7]). 为了提升主题建模在不同数据集(如短文本)中的性能, 研究人员开发了以 LDA 和 NMF 为基础的衍生算法[8-9]. 然而, 通过词袋表示的这些模型忽略了单词之间的语义关系,

词袋输入可能无法准确地表示文档.

近年来, 随着神经网络技术的兴起, 已有不少神经网络主题建模技术, 如结合变分自动推理(variational auto-encoder, VAE)[10]、对抗神经网络[11]、图神经网络(graph neural network, GNN)[12]等方法[13-16]的主题模型. 与词袋模型相比, 基于神经网络的主题建模更好地利用了词语间的语义相似度, 生成的主题一致性更高; 且神经网络主题建模的泛化能力更强[17]. 随着预训练的基于Transformer[18]的语言模型的发展, 能够得到更精确的上下文嵌入. 文献[19-20]分类讨论了不同主题建模技术的特性与局限性. Thompson 等[21]的研究表明, 通过对各种预训练模型生成的上下文词嵌入进行聚类得到的簇, 可以被解释为主题. 基于该结论, Angelov[22]提出 Top2vec, 通过 Doc2vec[23]将文档和单词联合嵌入后使用聚类技术得到主题, 并以簇的质心作为主题向量. 然而, Top2vec 的聚类基于密度, 而采词基于距离, 容易误采集其他簇的单词而导致建模质量下降. 随后, BERTopic[5]采用 c-TF-IDF 变体解决了 Top2vec 存在的聚类和采词不一致的问题, 并获得更好的可解释性、更高的主题一致性和主题多样性. 但是, 尽管神经网络主题建模技术有效地提升了建模质量, 仍然不能避免地产生低质量的主题. 因此, 需要分析这些主题的成因, 对其进行优化和改进, 以得到更高质量的结果.

1.2 主题建模的可视分析

基于主题建模的可视化为用户提供了直观的视图, 可以促进用户有效地解释与分析主题建模的结果. 迄今为止, 面向文档集合的内容分析和探索, 人们开发了各种交互式可视化方法. 如 Jiang 等[24]为多模态模型设计了多层可交互可视化分析界面; UTOPIAN[25]将半监督 NMF 与可视化结合, 设计了 *t*-SNE[26]降维散点图可视化文档分布, 并提供了主题-单词交互功能; LDAexplore[27]和 TMvis[28]等将 LDA 模型和可视化方法结合, 通过视图分析和用户交互探索语料库, 迭代模型得到新结果. Zhu 等[29]以加权方法优化控制 LDA 的计算过程, 并将改进后的 LDA 与可视化结合, 设计了一个基于一致性度量可通过前端界面编辑操纵的主题建模; Dou 等[30]提出了将时间导向加入视觉文本分析的挑战, 文献[31-34]在文本可视化中加入了时序要素, 设计了随时间发展的媒体文本的可视化分析; Poux-Médard 等[35]利用实体间的时间距离进行建模, 对不同数据集间的交互进行观测和展示, 并将其与可视化结合从而更直白地展现结果. 文献[36-38]采用层次结构来适应词汇表和文档数量的增加, 并且 Yang 等[39]设计了对层次聚类的交互式操作指导. Kowsari 等[40]将深度学习扩展到层次分类文档问题中, 通过多种深度神经网络, 即深度神经网络、递归神经网络和卷积神经网络相结合, 设计了文本层次深度学习架构 HDLTex, 并与传统的支持向量机模型和组合中各个模型进行消融实验, 证明了 HDLTex 模型的优越性; Zhang 等[41]以双曲 GNN 为基础设计双曲图主题建模网络, 将文档之间连接的网络层次结构与文本的语义层次结构集成, 以可视化的方式解释了其方法的有效性. 但是, 这些工作仅限于展示主题结构, 用户难以通过它们去快速地修改局部的主题, 提高主题质量. 在此基础上, 许多工作致力于人在回路上的主题建模, 使用户在可视化中参与和引导主题的修改. El-Assady 等[42]设计的系统允许用户评判使用不同参数生成的 2 个主题的优劣, 并使用强化学习迭代地提高模型质量; El-Assady 等[43]提出一种用户干预的可视化方法, 用户在浏览每个策略的预期结果后, 从几种交互策略中选择出最合适的一种, 但该方法在大型数据集上效率过低; Semantic concept spaces[44]结合语义信息指导主题模型, 并允许用户修正结果; Kim 等[45]设计了一种半监督层次聚类 NMF 模型 Architext, 并基于主题、文档和关键词 3 种交互层次设计局部交互算法, 引导用户局部地改进模型结果, 但其并未分析模型的结果, 用户只能根据自己的主观感知评判模型结果; 刘陈一等[46]支持用户选择表征后重构属性子空间, 但没有给用户操作建议, 需要用户自身有一定领域知识. Kuang 等[47]为 LDA 模型设计了一种基于主题一致性的修改方法 TopicRefiner, 通过主题一致性的变化指导用户改进 LDA 模型的结果. 对于一直以来产生的大量研究成果, 也有不少综述对其进行了整理与总结, 巫英才等[48]对文本可视分析及典型的文本分析和可视化技术进行了阐述与展望; Liu 等[49]对文本挖掘可视化的研究趋势及之间的联系进行了调查研究; 叶力等[50]对叙事逻辑的挖掘可视化分析进行了调查研究.

上述可视化方法大多基于词袋模型, 缺少面向深度神经网络主题建模的可视化系统. BERTopic 虽然提供了一些简单的基于 Python 的视图以帮助用户理解建模结果, 但用户无法通过交互地修改快速地提升模型的质量. 因此, 本文设计并开发了 BTVis, 以 BERTopic 模型为基础, 提供了面向非专业人员的模型理解和交互探索功能, 允许用户进

行局部交互, 引导用户修改模型的性能.

2　基于 BERTopic 的数据处理

2.1　BERTopic 模型

BERTopic 是基于深度学习的一种主题建模方法, 采用模块化的步骤, 允许用户自主地选择每个模块的算法. BERTopic 的主要建模步骤如下:

Step1. 文档嵌入. 使用预训练后的语言模型将句子和段落转换为向量表示, 从而得到文档嵌入的向量表示.

Step2. 降维与聚类. 基于向量表示的文档通过降维和聚类得到表示主题的文档簇.

Step3. 主题表示. BERTopic 使用基于类的 TF-IDF, 即 c-TF-IDF 得到每篇文档聚类的主题-词分布.

BERTopic 的参数包含 $n-\text{gram}$ (在主题表示中是否可以使用词组), min_topic_size (每个主题的最小文档数量)和 nr_topics (最终生成的最大主题数量); 此外, BERTopic 还可以选择不同的降维和聚类方法来构建主题簇. 其中, $n-\text{gram}$ 表示是否将词组视为关键词, 该参数范围值为(1,1)~(1,4), 数对左侧表示词组内单词最小数量, 右侧表示词组内单词最大数量; min_topic_size 与 HDBSCAN 的簇最小值相同, 该参数最小为 10, 最大可自行设置, 通常, 需要根据数据集的大小、数据密度和用户需求综合考量并测试来确定其值; nr_topics 决定模型最终生成的最大主题数量, 与 min_topic_size 共同影响模型的结果, 当生成的主题数量大于 nr_topics 时, 系统会根据相似度依次合并最相似的主题. 例如, 可以尝试 min_topic_size =10 和 nr_topics =20 来控制主题规模. 但是, 当 nr_topics 比实际生成的主题数少很多时, BERTopic 会不断地将主题两两合并; 由于文档数量的增加和语义上的混淆, 该主题将逐渐中心化, 获得与其他主题更高的相似性, 从而每次都参与合并, 导致最后出现一个完全无意义的大主题. 因此在实际使用中, 本文在 HDBSCAN 聚类的基础上设计了局部编辑方法替代 nr_topics, 以增强用户修改主题结果的可靠性, 但当 min_topic_size 过大或过小时, 主题效果也不佳. 为此, 本文设计了交互调整 BERTopic 参数的界面, 让用户通过调整该参数得到较好的主题生成结果. 除了 UMAP 降维方法[51]和 HDBSCAN, BTVis 还提供 *t*-SNE, 主成分分析降维方法和 *K*-means 聚类方法供用户选择, 且如果选择 *K*-means, 则只有单层扁平聚类结果. 在默认情况下, 使用 UMAP 降维方法和 HDBSCAN 层次聚类方法.

2.2　HDBSCAN 聚类

HDBSCAN 是一种层次聚类方法, 被应用于 BERTopic. 获得所有文档的嵌入表示后, 该方法首先计算所有文档之间的相互可达距离, 公式为

$$d_{\text{mreach}-k}(a,b)=\max\left\{c_k(a),c_k(b),d(a,b)\right\}.$$

其中, $c_k(a)$ 表示文档 a 与其第 k 个近邻的距离; $d(a,b)$ 表示 a 和 b 间的欧几里得距离.

得到文档间的相互可达距离后, 构建最小生成二叉树, 以获得树状层次结构. 其中, 每个叶节点代表一篇文档, 每个父节点代表其子节点中的文档之和. 由于二叉树的规模庞大, 需要将复杂的层次结构压缩成一棵较小的树. 从根节点开始向下压缩, 如果树节点的 2 个子节点内的文档数量均大于设定的最小文档数, 则保留并视为有效节点; 否则, 舍弃小于最小文档数的节点, 该节点内的文档被称为离群文档.

在压缩后的树结构中, 通过自下而上地计算每个节点的稳定性, 选择树节点作为最终聚类所得的主题. 节点稳定性的公式为 $\sum_{p\in n}\left(\lambda_p-\lambda_{\text{birth}}\right)$. 其中, n 表示节点; p 表示节点中的一篇文档; λ_p 表示文档 p 离开该节点时该节点的稳定值; λ_{birth} 表示父节点分裂生成该节点时的稳定值. 如果父节点的稳定性小于子节点的稳定性之和, 则选中这些子节点; 否则, 选中父节点作为最终聚类, 并取消选择它的子节点. 到达根节点后, 所有选择的节点作为最终生成的平面聚类结果, 每个节点作为一个主题.

根据 HDBSCAN 算法, 本文在获取压缩后的树结构的基础上, 收集每个树节点进行分裂时产生的离群文档, 将这些文档合并成一个节点, 并将该节点作为父节点的第 3 个子节点, 称为“离群节点”. 在树结构中, 最终被选定为主题的节点称为“主题节点”, 其他节点称为“过渡节点”. 将过渡节点和离群节点同样视为主题, 并计算各项指标, 辅助用户分析.

2.3　BERTopic 引入局部主题编辑算法

BERTopic 计算出初始主题后, 本文设计了主题、文档和单词 3 种不同的交互类型, 加强通过不同层次修改模型结果的可靠性.

2.3.1 主题级交互

主题级交互将主题内的绝大部分文档选入并重新定位至新的主题, 适用于用户修改较大范围的局部主题, 可以大幅地改变建模的结果. 当用户想进行小范围和更精确的修改时, 需要使用文档级交互和单词级交互. 主题级交互功能如下:

(1) 主题合并. 在 HDBSCAN 层次聚类结构中选择某一个父主题, 将其所有子主题合并至该主题中. 该方法将离群节点内的文档同样合并入主题.

(2) 主题分割. 选择一个主题节点, 将该主题内的所有文档视为新语料库, 用户可以选择新的最小文档参数运行 HDBSCAN. 将 HDBSCAN 根节点下的子节点作为分割后的 2 个主题, 在根节点分裂中失去的离群点被聚合为离群子节点. 该功能实现对主题的局部分割, 而无需对整个模型进行计算.

(3) 主题移动. 选择主题 A 和主题 B, 将主题 B 的所有文档置入主题 A 中再进行主题分割, 重新分配文档归属.

(4) 主题删除. 选择一个主题节点, 将该主题内的所有文档分配至兄弟离群节点内, 再删除该主题.

2.3.2 文档级交互

文档级交互通过用户观察并筛选主题内的文档, 使主题与所选文档更相关. 由于主题内文档的数量较多, 用户不可能花费大量的时间逐个查看文档内容来修改单个主题, 因此交互算法将与所选文档相似的文档加入计算. 由于 BERTopic 本身并不涉及通过文档层面修改主题建模, 因此本文获取 BERTopic 建模结果中的部分指标, 并设计了文档级交互条件. 文档级交互功能如下:

(1) 文档分割. 用户选中多篇文档后, 当该主题文档满足条件 $D_{\min}<\alpha d_{\text{Tscore}}+(1-\alpha)d_{\text{Wscore}}<D_{\max}$ 时, 将被分配至新的子主题, 剩余的文档被分配至另一个子主题. 其中, d_{Tscore} 表示所选文档的平均文档-主题相关性得分, 由文档与主题的语义向量的余弦相似度计算得到; d_{Wscore} 表示所选文档的平均文档-关键词得分, 由文档内关键词数量与文档长度加权得到; α 用于调节 2 种得分占比的参数, D 表示阈值范围, 且 α 和 D 均可调节.

(2) 文档移动. 对于用户选中的多篇文档, 当源主题文档满足上述条件时, 将选中的这些文档分配至目标主题的文档.

(3) 文档删除. 对于用户选中的多篇文档, 当源主题文档满足上述条件时, 将这些文档分配至兄弟离群节点中.

2.3.3 单词级交互

单词级交互通过用户观察并筛选主题内的关键词, 使主题与所选关键词更相关. 由于 BERTopic 中的关键词并未参与建模结果的生成, 仅用于描述主题, 因此设计了单词级交互条件. 单词级交互功能如下:

(1) 单词分割. 对于用户选中的多个单词, 将该主题下的每篇文档单独视为一个类, 使用 c-TF-IDF 得到每个主题的关键词, 即每篇文档的关键词. 当文档关键词与用户所选关键词相同时, 计算该关键词的权重; 若权重高于阈值, 则分配至新子主题, 剩余的文档被分配至另一个子主题.

(2) 单词移动. 对于用户选中的多个单词, 当源主题文档符合单词分割中的上述条件时, 选中这些文档分配至目标主题的文档.

(3) 单词删除. 对于用户选中的多个单词, 当源主题文档满足上述条件时, 将这些文档分配至兄弟离群节点中.

(4) 单词停用. 对于用户选中的多个单词, 将这些单词置入全局停用词, 全局停用词中的单词将不会再出现于主题中.

2.4 主题标题自动生成

已有的研究通过一些高权重的单词传达主题, 要求用户自行将这些单词联系起来并形成整体印象. 但是, 当用户需要快速浏览主题时, 仅展示几个单词可能会增加认知负担. 为了解决这一问题, 本文使用国内千帆 ModelBuilder 平台的 ERNIE-Speed-128K 模型自动为主题生成标题. 向 ERNIE-Speed-128K 模型提供主题单词、词权重和 3 篇代表性文章, 以便该模型能够输出不超过 5 个词的句子, 用于概括主题内容, 帮助用户更快地获取信息. 当主题质量较高时, 生成的主题标题更具代表性; 否则, 生成的主题标题容易出现明显错误, 可以帮助用户发现低质量的主题.

3 BTVis

3.1 系统目标

通过与目标用户, 即对文档语料库的主题内容感兴趣的一般用户进行讨论, 确定用户需求和的设计目标. 让用户查看 BERTopic 模型直接生成的结果与视图, 并记录他们的使用情况, 以及遇到的问题和困难.

(1) 用户指出，这些静态的可视化视图功能单一并且视图较分散，导致分析的时间成本大幅增加，设计一个更优化的、支持对 BERTopic 建模结果全面分析的可视分析系统是很有必要的.

(2) 因为用户不能具体地查看单个主题内的文档内容，所以也不能评判文档与主题之间的关系；同时，缺乏辅助手段帮助评判主题结果的优劣，难以比较主题的结果是否冗余.

(3) 用户无法使用模型提供的参数精确地修改或优化主题的内容，模型参数的调整与主题调整无法直观对应. 希望能有更加直观和方便的修改方式，帮助精确优化主题结果.

因此，在充分考虑用户的意见和分析 BERTopic 模型存在的问题的基础上, BTVis 的设计目标如下:

G1. 展示 BERTopic 建模中的层次聚类过程. 基于 HDBSCAN 的层次聚类过程提供可视化的展示，揭示主题的分类过程和层次结构信息，帮助用户理解 BERTopic 生成的建模结果.

G2. 挖掘离群文档，发现 BERTopic 建模结果中潜在的问题. 按照层次结构对离群文档进行分类，对分类后的离群文档同样进行主题建模，结合主题建模得到的信息量化指标和可视化技术，更加深入地挖掘主题、文档和单词 3 个层次的内容，辅助用户更深入地评估主题的质量，引导用户纠正部分离群文档的所属主题，最终得到质量更高的结果.

G3. 用户通过局部编辑渐进式地优化主题. 通常，修改模型结果时需要通过大量的调参，而较小的调整也会导致不可预测的偏差. 当用户发现问题并尝试修改主题时，需要结合模型提供交互功能，修改局部主题，将变更的结果缩小至单一主题或部分主题的范围内. 因此，需要设计主题级、文档级和单词级 3 种层次的交互算法，允许用户通过不同细节的方式帮助修改局部范围的主题.

G4. 实现多视图联动和交互探索的 BTVis. 通过交互探索，用户可以逐步地探索和挖掘感兴趣的主题；通过视图联动，为用户同时展示主题、文档、单词和层次结构的整体信息，增强用户对模型的理解且方便对主题结果进行评判.

3.2　工作流程

BTVis 由数据处理和可视分析系统 2 部分组成，其流程如图 1 所示.

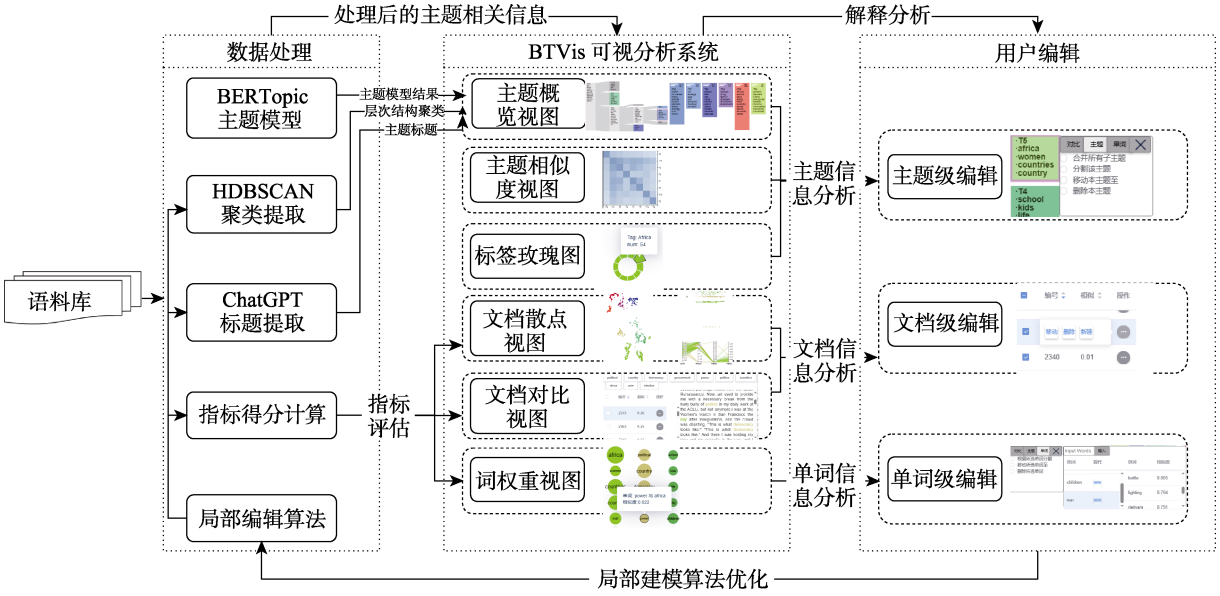


图 1　BTVis 整体流程

在数据处理部分，首先对数据预处理得到语料库，运行 BERTopic 获得主题建模结果；然后提取 HDBSCAN 聚类过程获取模型结果生成过程的信息；再计算主题、文档和单词的多种指标得分，辅助评估主题模型结果；最后调用 ChatGPT，根据主题的文档和单词为各主题生成标题. 通过设计局部建模算法，快速地提升生成的局部模型结果.

对数据进行处理并提取结果后，将结果传递至可视分析系统中，使用户可以通过主题概览视图、主题相似度视图、标签堆叠图、文档散点视图、文档对比视图和词权重视图进行多粒度的探索. 在可视分析系统中，主题概览视图支持探索主题级别的层次结构信息，主题相似度视图和标签堆叠图辅助用户定位和比对主题信息，文档散点视

图和文档对比视图支持更细粒度的探索和比较不同主题内部的文档信息，词权重视图支持比对主题间的关键词信息. 用户根据多层次信息发现潜在的问题后，通过主题级、文档级和单词级编辑功能调用局部编辑算法. 重新处理相应的主题建模结果数据后，可视分析系统将更新局部建模的结果，实现引导和优化. 通过直观的交互反馈，用户可以循环进行探索与交互，提升模型质量并得到可信任的主题.

3.3 设计需求

为了实现第 3.1 节提出的系统目标，提出设计需求如下.

R1. 提供 BERTopic 主题建模结果的概览与细节信息(G1, G2, G4). 用户可以获得全局信息，自主地交互探索层次结构信息，查看感兴趣的主题节点、过渡节点和离群节点.

R2. 提供查找相似主题的方法(G1, G2). 当用户确定感兴趣的主题后，可以快速地查找相似的主题，以节约时间.

R3. 提供主题内文档的语义信息，比较不同主题的属性间的异同(G1, G2, G4). 用户通过文档的语义信息和主题的属性，定位主题内存在异常的文档与单词，增强用户对主题的理解.

R4. 提供多种辅助判断建模结果优劣的方法(G2). 用户可结合定量指标、辅助标签、高权值文档等多种方法，分析主题的优劣并判断异常出现的原因.

R5. 提供优化模型结果的编辑功能(G3). 用户根据对主题的理解，编辑优化主题至理想状态.

3.4 视图总览

本文设计并实现了如图 2 所示的主题建模可视分析系统. 为了确保用户便捷地搜索感兴趣的主题，通过自上而下的方式展示主题层次结构，并按用户需求展示更深层次的主题信息. BTVis 由 6 个部分组成，参数窗口可以调整 BERTopic 模型选择的降维和聚类方法，通过选择后点击重新运行，可得到初始化建模结果. (1) 主题概览图(R1). 提供 HDBSCAN 层次聚类结构和最终生成的主题的概览，用户可以通过探索并定位至感兴趣的主题，使用交互面板在其他视图中展示主题的细节信息，当用户点击一个主题时，其他视图会展示该主题相关的多粒度信息. (2) 主题相似度视图(R2, R3). 比较同层次主题间的相似性信息. (3) 文档散点图(R2, R3). 给出不同主题的文档分布情况和每篇文档的属性，帮助快速地查找和定位存在问题的主题文档. (4) 文档概览视图(R4). 根据选择的主题和文档散点，展示相应文档的文本信息和单词分布，提供不同主题下不同文档间的对比功能. (5) 标签图(R4). 展示所有主题或所选主题下文档的标签信息，辅助用户评价主题的可靠性. (6) 词权重视图(R4). 展示选定主题间的关键词的权重信息和语义信息，评估主题结构信息的混淆度. 汇总这些信息后，用户可以通过编辑功能(R5)优化局部模型结果.

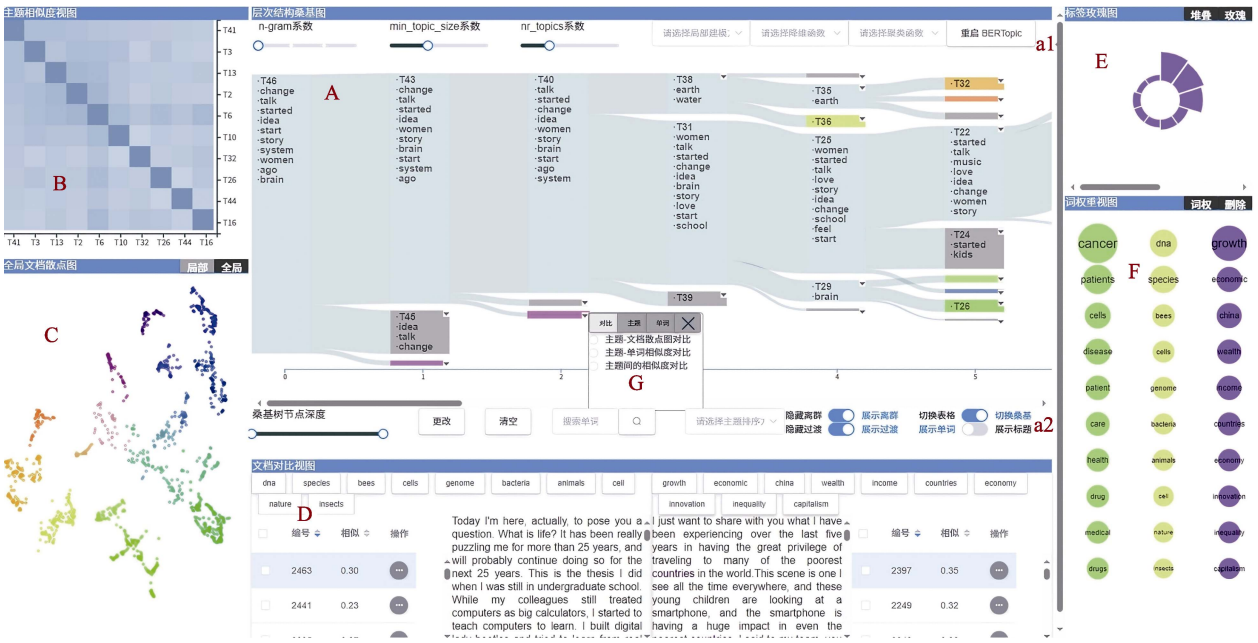


图 2 BTVis 界面

3.5　主题概览图

主题概览图如图 2 中的 A 部分所示. 根据需求 R1, 层次主题概览图提供桑基树和表格图 2 种类型的层次可视化视图.

图 2 中的 a1 部分是 BERTopic 的全局参数, 包括 $n-\text{gram}$, min_topic_size, nr_topics, 建模方法, 降维函数和聚类函数.

用户可以调节并点击重运行按钮, 重新生成主题建模结果. 桑基树中提供 HDBSCAN 对文档聚类生成的层次结构(R1), 每个矩形代表一个节点, 矩形的高度编码主题内文档的数量; 主题节点的矩形的颜色与文档散点视图的颜色对应, 离群节点的矩形固定为深灰色, 过渡节点的矩形固定为浅灰色; 矩形内的单词代表权重从高到低的关键词概览, 对非主题节点的矩形也计算了它们的各项属性. 对于矩形高度不足以展示多余矩形高度的数量的关键词的节点, 可以点击下拉按钮显示前 10 个关键词. 用户在“搜索单词”输入框输入单词后检索拥有该关键词的主题, 并将其高亮显示. 通过图 2 中的 a2 部分的桑基树节点深度条和更改按钮调整展示的层级, 或双击某个主题, 将该主题定位至视图最左侧. 单击主题时, 文档散点图会高亮显示当前主题的所有文档, 且标签堆叠图展示当前主题内文档的标签总览. 若该主题存在未展示的子级别主题, 则展示子主题. 通过图 2 中的 a2 部分的过渡按钮和离群按钮分别显示和隐藏非主题节点的关键词和离群节点.

通过切换按钮, 可以将当前桑基树中的主题节点放置于表格图, 如图 3 所示. 在表格图中, 矩形上方的白线和数值代表该主题的定量分析指标、主题一致性和主题多样性[52]. 主题一致性由归一化点态互信息(normalized pointwise mutual information, NPMI)[53]计算, 主题多样性即主题唯一词的百分比. 主题下的文字展示当前主题的标题或关键词, 用户可以使用图 2 中的 a2 部分的展示单词或标题按钮切换标题和关键词. 矩形的长度代表主题内的文档数量, 点击下拉按钮时将会拉伸主题的长度, 以显示更多的关键词. 用户可以通过选项框选择文档数量、主题一致性、主题多样性和索引对主题进行排序, 快速地查阅相关主题. 用户可以自由选择并对照 2 种视图.

右击主题时, 如图 2 中的 G 部分所示, 在主题右侧展示交互面板(R5). 交互功能分为视图联动展示和编辑功能 2 类. 视图联动展示中, 主要功能



图 3　主题结果表格图

有主题-文档散点图对比、主题-单词相似度对比和主题相似度对比 3 种; 主题编辑功能中包括主题级编辑与单词级编辑. 单击主题按钮中的单选框, 即可进行主题编辑交互; 单击单词按钮, 输入需要交互的单词, 则候选栏推荐与用户所选单词的前 10 位最相似的单词, 方便用户筛选. 在可视化界面上, 以单击的形式完成各类编辑功能, 用户可以便捷地修改主题结果.

3.6　主题相似度视图

主题相似度视图如图 2 中的 B 部分所示, 展示主题之间的相似度与向量分布(R2). 当用户选择一个主题时, 系统自动地计算其他主题的余弦相似度并创建一个相似度矩阵. 为了调整主题规模与用户感官之间的平衡, 只选取与当前主题最相关的前 10 个主题进行排序, 帮助用户快速地选择相似的主题.

3.7　文档散点图

文档散点视图分为全局和局部 2 种模式. 全局模式如图 2 中的 C 部分所示, 展示降维至二维后的文档向量分布. 每个圆节点编码一篇文档, 节点在空间中的位置编码空间分布, 节点的颜色参考文献[54]将文档的向量降维至二维后映射到 colormap.

局部模式由散点图和平行坐标图组成, 用户可在交互面板点击主题-文档散点图对比按钮, 展示该主题的信息. 如图 4 所示, 圆节点编码文档; 节点的纵坐标编码文档嵌入向量的一维降维分布, 节点的横坐标编码文档与该主题的相关性分数; 圆的大小编码关键词密度; 圆上的弧长编码关键词得分. 通过平行坐标图, 用户可以调整上述 4 个属性的范围以筛选节点. 点击任意一个节点, 可在文档概览视图中展示相关主题和当前文档的内容. 该视图提供了多维度的文档指标, 帮助用户更快和更准确地筛选出文档在主题中的分布情况, 更细粒度地评估主题分类的性能.

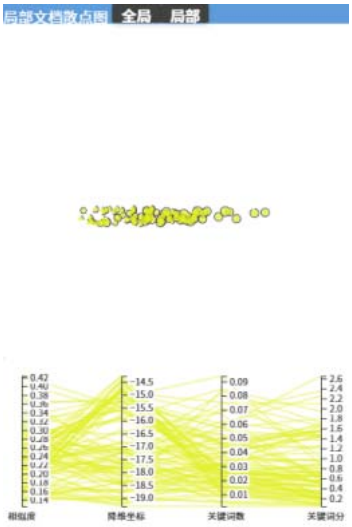


图 4 局部散点图示例

3.8 文档概览视图

文档概览视图如图 2 中的 D 部分所示，展示主题内文档的具体内容(R4). 受到文献[42]的启发，本文采取对比的方法，用户通过对比得出 2 个主题之间的差异和相似性. 文档概览视图按关键词权重排序展示 10 个主题关键词，并使用表格存储主题的所有文档，通过索引或相似性分数排序和搜索文档. 点击表格将展示当前的文档，文档中的关键词被赋予与其所属主题一致的颜色. 如果存在对比主题的关键词，则会赋予对比主题一致的颜色；若 2 个主题有相同的关键词，则加粗显示. 用户在层次概览视图或文档散点图中选择任意一个主题或文档时，文档概览视图将展示该主题并定位至所选文档. 用户可勾选文档，并通过按钮选择不同的文档级编辑操作(R5)，对建模结果进行修改.

3.9 标签堆叠图

标签堆叠视图首先以柱状图的形式显示生成的所有主题的标签，如果生成的主题过多，则下方会出现滚动条，便于查看所有主题的标签，如图 5 所示.

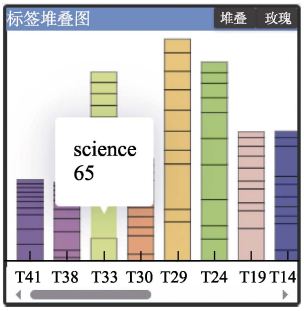


图 5 标签堆叠图

单击某个主题时，可以展示当前所选主题的标签分布信息(R4)，如图 2 中的 E 部分所示. 根据扇形的母线长度，可以显示标签数量占比关系；将鼠标放置在扇形上，则展示当前的标签和数量. 需要注意的是，根据使用的数据集不同，标签所代表的内容和意义有所不同. 本文只显示前 10 位数量的标签.

3.10 主题单词评估图

主题单词评估图用于分析不同主题间关键词的异同，如图 2 中的 F 部分所示. 使用 Word2Vec[55] 计算整个语料库中关键词的向量(R4). 用户点击“主题-单词相似度对比”按钮，交互展示当前所选主题的前 10 个关键词，圆的大小和顺序对单词在主题中的权重大小和重要性程度进行编码. 通过点击某个单词，可以计算其他单词和该单词的相似度，此时，圆的大小编码单词和所选单词的相似性程度. 通过权重和相似性，用户可得到主题的内容倾向，同时发现存在混淆分类的情况.

4 BTVis 评估

使用 TED 数据集作为实验数据集进行案例分析，验证 BTVis 的有效性.

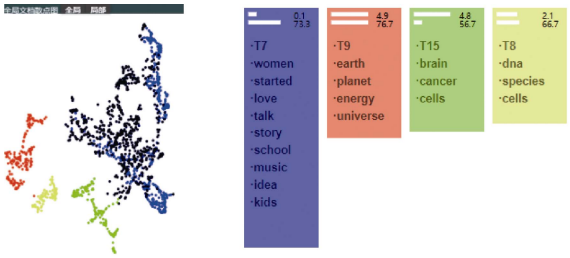
4.1 TED 数据集

TED 数据集是一个演讲内容数据集，其中的每篇文档都是一篇演讲稿，可视为长文本，其涵盖各个领域，包括技术、教育、政治和环境等. 该数据集共包括 2 470 篇演讲文档，每篇文档都拥有多标签，以标注该文档真实的主题范围. 对该数据集文档进行数据清洗和停词预处理，其中，玫瑰图展示的是一个主题内标签的数量.

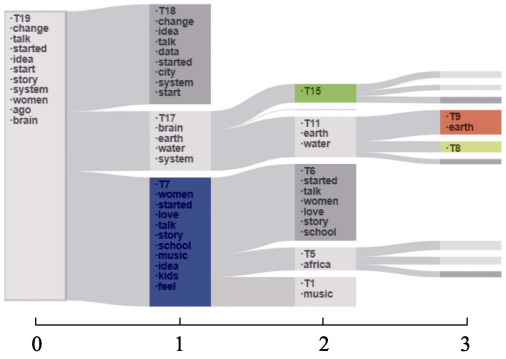
案例 1. 用户的目标是得到一个较好的主题建模结果，再探索主题间的关联性与差异性，并验证和解释效果较好的主题.

(1) 调节参数 *min_topic_size* 为 40，从图 6a 可以看出，该主题内文档的语义分布广泛. 生成的结果如图 6b 所示，共生成了 4 个主题. 逐个浏览发现，T7 主题文档数量大并且语义混淆. 用户点击图 6c 中该主题下的离群节点 T6，查看图 6a 中的黑色节点发现，该离群节点内的文档数量广泛，且语义离散，说明这些文档难以有效地聚类至 40 个以上，导致离群主题 T18 和 T6 中的文档数量过大，使得 BERTopic 模型生成的最终结果较差.

(2) 下调 *min_topic_size* 参数至 30 时，如图 7 所示. BERTopic 最终生成了 7 个主题，其中，T32



a. 散点图分布　　b. 主题信息

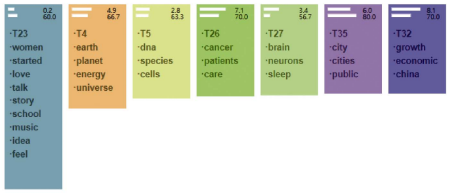


c. HDBSCAN 层次结构

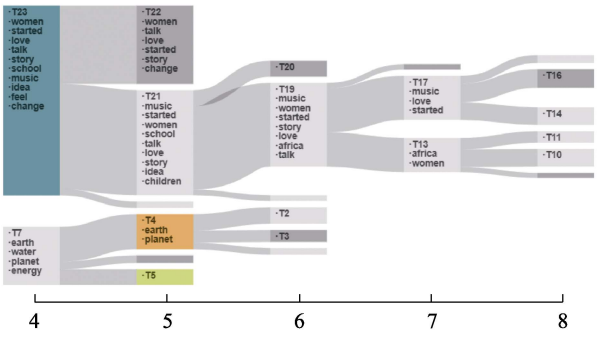
图 6　*min _ topic _ size* =40 下的主题结果概览



a. 散点图分布



b. 主题信息



c. HDBSCAN 层次结构

图 7　*min _ topic _ size* =30 下的主题结果概览

经济主题和 T35 城市主题是由图 6c 中的离群主题 T18 聚类得到，并且它们的主题一致性得分较高，说明随着 *min _ topic _ size* 的降低，BERTopic 能够逐渐得到拥有明确语义的主题；但是，从散点图分布和 HDBSCAN 的层次结构来看，主题 T23 与图 6c 中的 T7 基本相同，没有得到改善. 尽管主题 T23 后分裂生成更多的子主题，但是最终 BERTopic 并没有选择这些子主题，说明离群节点 T22 明显地影响了主题的稳定性，需要继续下调参数，挖掘离群节点中的其他主题.

(3) 将 *min _ topic _ size* 下调至 25 后，得到一个大致满意的结果，如图 2 所示. 可以看出，BERTopic 生成了 15 个主题，这些主题基本位于叶节点；对比图 7a 和图 2 中 C 区域可以看出，图 7a 中右侧的 T23 主题被有效地分割为 10 多个主题簇，相应地，每分成一个簇也产生了一定数量的离群点.

(4) 比较使用其他降维方法和聚类方法是否会产生更好的结果. 首先使用 *t*-SNE 降维方法，在 HDBSCAN 聚类方法下，该方法只生成了 4 个主题，其文档散点图如图 8 所示.

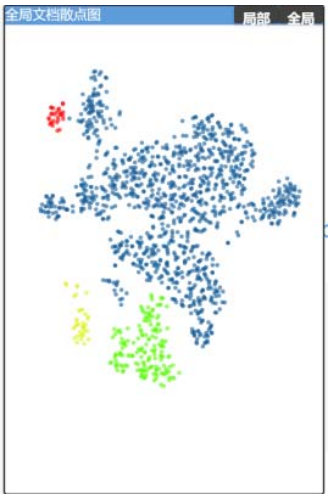


图 8　使用 *t*-SNE 降维方法在 HDBSCAN 聚类下的散点图

(5) 将 *min _ topic _ size* 参数下调至 10 发现，主题数量甚至降低为 2 个，如图 9 所示. 可以看出，HDBSCAN 无法对 *t*-SNE 降维下的 TED 数据集做出有效的分类.

(6) 将 HDBSCAN 替换为 *K*-means 聚类方法，并设置主题数为 15 个. 生成的散点图和主题结果分别如图 10 和图 11 所示. 可以看出，用户从主题单词出发，存在一部分主题完全相似，并且所有主题相似度都偏低.

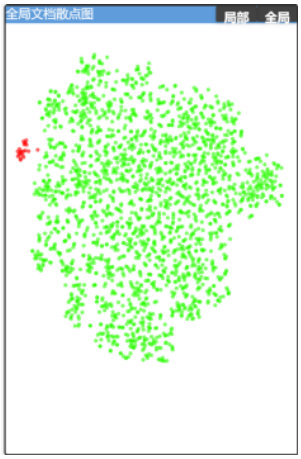


图 9 *min_topic_size* =10 时使用 *t*-SNE 降维方法在 HDBSCAN 聚类下的散点图



图 10 使用 *t*-SNE 降维方法在 *K*-means 聚类下的散点图



图 11 多样性排序下的主题结果

(7) 继续在 UMAP 降维方法下使用 *K*-means 聚类方法，生成的结果如图 12 所示. 从主题多样性指标的排序发现，绝大多数主题的多样性指标偏低，说明这些主题超过一半的关键词在其他主题中出现；在散点图中框选的部分，这些文档点分属于不同的主题，且 *K*-means 聚类过程并不删除离群节点，可能导致最终生成的关键词更加相似.

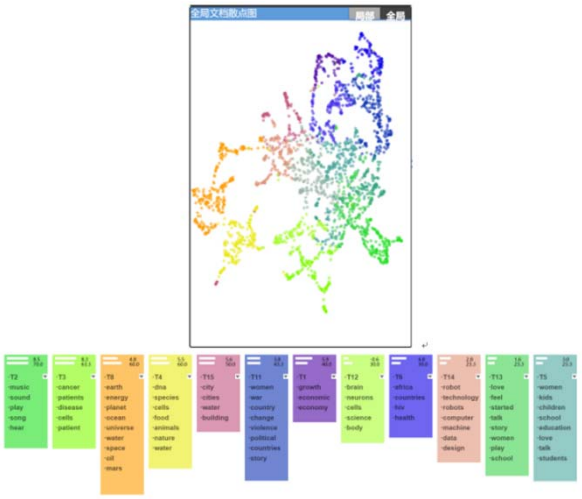


图 12 使用 UMAP 降维方法在 *K*-means 聚类方法下的主题结果

(8) 用户认为，UMAP+HDBSCAN 在聚类过程中剔除了一部分语义不相关的文档，能够增强主题的一致性与多样性指标，使主题内的语义更唯一，聚类更紧密，在 TED 数据集环境中更适配. 因此，在该环境中，BERTopic 模型使用 UMAP+ HDBSCAN 的初始结果在用户主观感知和定量指标比较上最好.

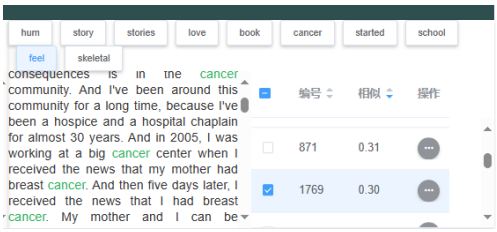
(9) 以 UMAP 和 HDBSCAN 生成的建模结果为基础，开始探索 BERTopic 生成的主题. 从图 2 中的 C 部分可以发现，左侧的橙黄色主题 T2，T3 与其他主题相对更偏僻，它们分别对应海洋环境主题和土地能源主题；此外，T3 也包含了一小部分描述土地环境的文档. 这些事实说明，T2 和 T3 中存在一部分使它们的语义更加接近的文档内容. 同样地，右侧的蓝色主题 T13，T21 和 T24 分别对应女性主题、战争冲突主题和教育主题. 在主题 T24 中包含 women 的关键词，在主题 T13 中有一部分文档也提及了 girl 和 women 的教育；说明女性作为一个独特的语义信息，在 TED 中被广泛地在不同领域中提及，同时也证明了这些语义相近的主题具有一定的内在内容联系.

(10) 准备优化质量较低的主题. 首先根据主题一致性对主题排序，并查看它们的标题. 主题一

致性最低的 T15 主题标题为“Empowering cancer survivors’ extraordinary stories”，涉及癌症患者述说自己的故事，如图 13a 中的第 2 个主题所示. 然而从图 3 中发现，主题 T9 中的单词与癌症和健康的语义更加接近. 因此用户将标题替换为主题词，发现主题 T15 中单词 cancer 的权重相对靠后. 之后快速地浏览该主题内的高权重文档，如图 13b 所示，发现仅有编号 1 769 的文档是演讲者描述自己与癌症抗争的故事，其他文档都是演讲者描述他们独特的故事. 于是，用户将这篇文档移动至 T9，在新的 T15 主题中，单词 cancer 消失，并且主题一致性提高了 0.9.



a. 主题 T15 变化



b. 主题信息

图 13　主题 T15 结果变化及相关文档

另一个主题一致性较低的主题 T18 并没有明确的语义，用户在桑基树上找到了该主题与其兄弟离群节点 T19，如图 14a 所示. 发现黑色的离群节点数量较多，基本分布在 T18，T15 和 T14 主题周围，且该离群节点 T19 中的单词同样没有明确的语义. 用户将 T18 主题置入局部散点图中，如图 4 所示，发现 T18 中的语义分布实际并不紧密. 用户根据平行坐标图，选择了关键词数和关键词分远超于其他主题的文档，发现该文档大量使用了单词 flag 和 design 描述旗帜设计，而其他的文档几乎没有涉及. 用户进一步确认了该主题语义不明确，不适合作为一个单独的主题，于是将 T18 主题删除，主题内的文档都移入 T19 中.

4.2　豆瓣影评数据集

豆瓣影评数据集是从豆瓣爬取的电影评论数



a. 离群散点分布　　b. 删除前的层次结构

c. 删除后的层次结构

图 14　删除主题 T18

据，每篇文档是一条不超过 100 词的评论，可视为短文本，共爬取了 50 000 个电影的 4 389 184 条评论数据. 在实验中，对文档进行数据清洗和停词预处理，并选择其中 100 个电影和 12 154 条相关评论作为实验数据，每条数据包含电影名和评论内容，每条电影的评论根据用户的打分比例抽取. 在文档嵌入阶段不进行停词处理，因为评论的文本量少，若进行停词，则在语义上会出现误差；在主题表示阶段使用停词，避免关键词出现无意义词. 该数据集中，玫瑰标签图展示的是一个主题内评论的电影名称和该电影影评的数量.

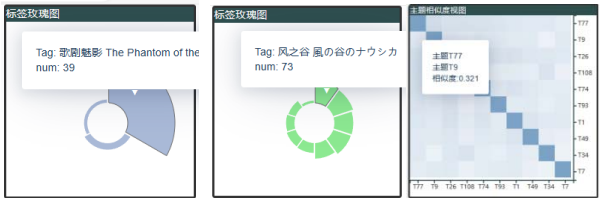
案例 2. 用户的目标是分析生成主题的好坏，并根据自己的理解优化主题结果. 经过调试后，将 *min _ topic _ size* 设置为 38. BERTopic 初始化共生成 47 个主题. 图 15 所示为其中一部分主题.



图 15　部分主题初始生成结果

(1) 初步分析这些聚合的主题是否有效. 切换至表格图，逐一查看表格内的标题和单词. 如图 16 所示，用户根据主题信息和玫瑰图发现，一部分主题与单个电影高度相关，如主题 T92 主要评论电影《歌剧魅影》中的音乐魅力；另一部分主题包含

多个不同电影的评论，如主题 T77 描述了宫崎骏的一部分动画电影：说明主题既可以分类一部电影中的主流评论，也可以分类多部电影中的相似评论．之后，查看主题相似度视图发现，大多数主题的最高相似度不超过 0.3，证明这些主题在内容上区分度高；此外，在该语料库中，主题一致性指标都相对较低，最高仅 5.7，最低为−31.7．



a. 主题 T92 玫瑰图　b. 主题 T77 玫瑰图　c. 主题 T77 相似度

图 16　不同类型主题的玫瑰图和主题相似度

(2) 使用主题多样性和主题一致性排序，查看指标较低的主题．首先发现一些主题中没有明显语义的词较多，如 T1 主题中主要是针对导演进行评价，但词语中包含了“看到”“知道”“看过”等没有明确语义倾向的词，于是将发现的这类单词加入停用词，如图 17a 所示．重新计算主题表示后，T1 主题的指标有一定的提升，但还存在一些无意义的词，于是使用删除单词功能，输入无意义的词和相似词，新生成的主题如图 17b 所示，使得 T1 主题更加清晰，该主题内的文档大多是对导演执导的文艺片持负面评价．



a. 停用所选单词后的 T1 主题单词分布



b. 删除所选单词后的 T1 主题单词分布

图 17　T1 进行停词和单词删除

(3) T36, T30 和 T7 这 3 个主题的头 2 个关键词均为“喜欢”“好看”，并且主题一致性都较低．使用词相似度视图和文档概览视图发现，主题 T36 的评论持负面意义，即“不喜欢”“不好看”；其他 2 个主题均为正面意义，且这 2 个主题的其他关键词高度相似，都是赞美演员或电影．于是将主题 T30 和主题 T7 主题合并，合并后的主题 T30 如图 18 所示．



图 18　T30 和 T36 相似主题比较与合并结果

(4) 主题 T33 尽管指标都相对较高，但是关键词和其他主题有明显区别，即主题内缺乏情感和评论性质的单词．因此，用户点击主题，查看了该主题的局部散点视图和部分文档．如图 19 所示，在局部散点图中绝大部分文档的主题相关性都低于 0.2，高相关的文档中，大部分是对电影相关信息的介绍式评论；低相关性的文档则夹杂了部分无意义英文、网址和数字．用户认为，该主题本身在分类上没有问题，但该主题并不包含明确语义，与一般的影评关系低．于是，用户决定删去这个主题．

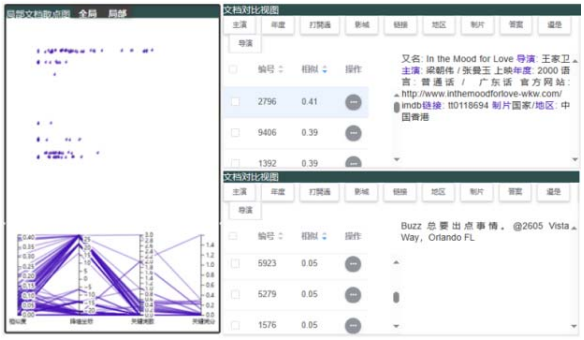
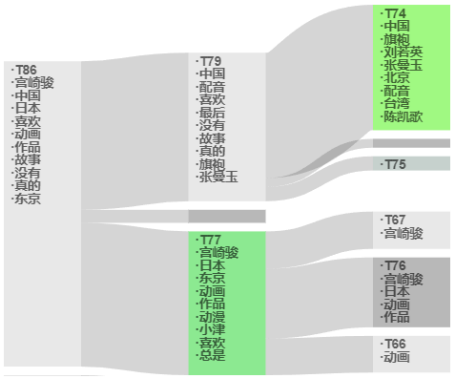


图 19　T33 文档信息概览

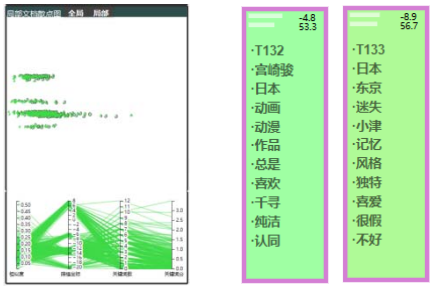
(5) 探索文档数量较大的主题．主题 T77 是描述宫崎骏动画电影的主题，但是关键词中出现了“东京”“小津”等词，用户发现这些词指向《东京物语》及其作者小津安二郎．进一步，使用局部散点图对主题进行分析后发现，该主题的文档在降维坐标上呈现 4 种不同的维度．通过平行坐标图筛选评论后发现，降维坐标为 6 左右的评论主要在讨论《东京物语》电影和其他非宫崎骏导演的日本电影，降维坐标为−7 左右的评论混杂了《东京物语》

与宫崎骏动画，而剩余 2 个维度的评论主要在讨论宫崎骏与日本动画.

用户从如图 20a 所示的桑基树中发现，该节点在 HDBSCAN 层次聚类中仍可后续生成 2 个主题，但是主题并没有很好地分类，并且出现了大量的离群点，其关键词仍为“宫崎骏”“日本”“动画”，说明在该主题下，设置参数 min_topic_size=38 并不能很好地将小津安二郎和宫崎骏相关作品的影评区分开. 对 T77 主题进行主题分割，设置参数 min_topic_size=20. 新生成的主题 T132 和 T133 中，T133 只有 22 篇文档，但包括了“日本”“东京”等词. 为了继续将相关文档从 T132 移出，用户使用文档移动和单词移动将相关文档扩充至主题 T133. 移动后的结果如图 20 所示.



a. 主题 T77 层次聚类结构



b. 主题 T77 散点图　　c. 主题 T77 分割最终结果

图 20　T77 分割结果

用户结束探索后，通过 BTVis 的各个视图，能够理解通过调整参数 BERTopic 如何生成主题. 得到良好的初始结果后，能发现主题建模中潜在的问题，并从辅助视图中挖掘隐藏的信息，帮助用户解释出现这种情况的原因；同时，用户能够使用编辑功能优化其中的错误，得到更加准确的分类，证明了该系统的有效性. 在实际应用中，用户并不能一次性地得到高质量的结果，在探索中可能也会降低模型的质量，需要通过多次探索和交互选择以利于获得高质量的模型，形成渐进性的过程.

4.3　用户实验

通过用户实验，验证 BTVis 各项功能的实用性. 实验中，招募了 100 名参与者，年龄为 20~25 岁，包括 79 名计算机系研究生和 21 名非专业人员. 其中，部分参与者拥有可视化方面的经验，但并未了解过主题建模，其余参与者则没有可视化方面的经验.

实验中，使用 BTVis 作为工具，评估 BERTopic 模型解释的质量. 参与者首先花费 10 min，依据系统演示视频和系统操作手册了解系统中各个视图的内容，并要求在 20 min 内完成指定的任务. 用户需要在完成每个任务后给出任务结果，并对部分问题在 0~5 之间打分. 在 TED 数据集和大运河数据集上，用户得到的初始主题建模相同.

大运河数据集搜集关于大运河的资料，每篇文章都可视为长文本，包括发表的时间. 大运河数据集共有 1 723 篇文章，文章的发表时间是其标签.

用户实验的任务如下:

(1) 结合关键词和辅助视图，概览主题在 HDBSCAN 聚类中的层次结构，查阅主题标题和主题单词，选择 3 个主题，简要描述它们的内容;

(2) 查找关键词，查看主题的层次流向，并评价这些主题之间的关联性和易见性;

(3) 选定主题并查看相似度视图、评价相似度最高的主题，以及所选主题在桑基树上是否接近;3 个较接近的主题的关键词相似度是否高，标签是否接近.

(4) 结合离群节点、辅助视图和定量指标查找有待改进的主题，使用交互操作编辑主题，判断修改后的主题是否改进，并记录修改前后的数据变化.

上述任务中，任务 1 用于测试可视化系统能否有效地表达主题信息，任务 2 用于查看 HDBSCAN 层次聚类结构所得主题的关联性，任务 3 查看相似主题之间的关联，任务 4 测试系统对主题建模结果的可解释性并评估主题编辑功能的效果.

任务 1 中，用户基本可以通过主题单词总结出主题大致内容. 在大运河数据集中，用户总结了如 T16 数字传媒、T20 大运河的世界文化遗产等；在 TED 数据集中，用户总结了如 T1 太空科学、T23 机器人与人类、T37 脑科学等: 说明大部分主题的关键词基本上能有效地表达主题的内容；但是，也存在部分主题因为关键词语义不清晰，所以用户没有总结的情况，如 TED 数据集中的 T15 主题.

任务 2 中，用户查询了如 music, human, 文化等关键词，并清晰地看出查询所得主题间的关联

性，关键词从某一深度上的过渡主题开始出现，并最终流向底层主题.

任务 3 中，用户基本认为相似度高的主题在桑基树上也是靠近的，同时，相似度高的主题之间关键词相似并且标签接近，如图 21 和图 22 所示；说明桑基树的层次结构合理有效，生成的主题可用度高. 对于部分主题不靠近的情况，在 TED 数据集中不同文档生成的主题之间基本独立，和其相关的主题相关性均小于 0.55.

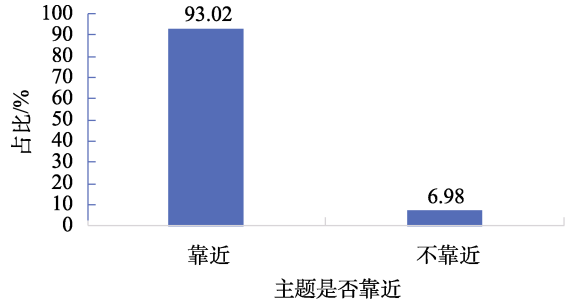


图 21 主题是否靠近统计

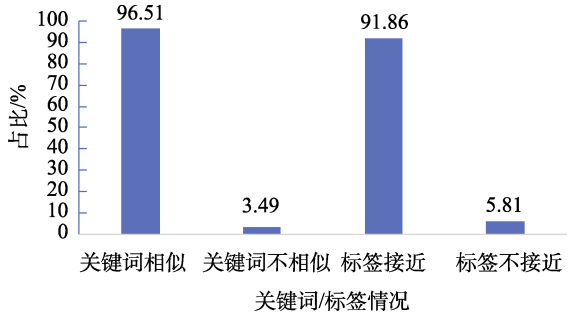


图 22 关键词相似度与标签是否接近统计

任务 4 中，所有用户修改主题的统计次数如表 1 所示. 可以看出，多数用户认为 TED 数据集上的主题 T37 和 T15 需要改进，说明用户能够依据辅助视图找出可以优化的主题及其存在的问题，也有部分用户结合辅助视图后尝试改进其他主题. 大部分用户认为，通过辅助视图可以发现需要改进的主题，并能够进一步找出主题内存在的问题. 当用户对它们预想的主题进行编辑操作后，本文统计了用户主题改进后平均连贯度指标的变化，结果如图 23 所示. 可以看出，当用户正确地优化主题后，主题一致性得到一定的提升；若用户进行了错误操作，也会导致指标下降.

表 1 用户认为需要修改最多的 5 个主题

排名	关键词	频次
1	T37	11
2	T29	8
3	T28	8
4	T15	7
5	T27	7

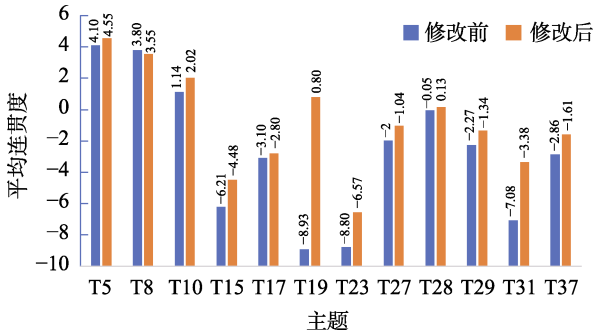


图 23 主题指标结果统计

对用户实验的任务进行 NASA-TLX 实验[56]，结果如表 2 所示. NASA-TLX 指数是一种通过精神需求、物理需求、时间需求、表现、努力和挫折 6 个尺度来评估一项工作的综合负荷的多维评估方法，得分为 1~5 分，得分越高，表示在该尺度上的工作负荷越低，表现越好. 可以看出，平均分值在 4 分左右，说明 BTVis 具有可操作性；有 16 个用户在物理需求和挫折尺度给出 1~2 分，通过与他们进行交流发现，其中 10 人是非计算机专业用户，在使用系统初期，他们难以轻易上手操作，在熟悉系统使用时花费了更长的时间；也有用户在使用交互操作后产生的结果与预期不符，指标并未得到提升.

表 2 NASA-TLX 实验结果

尺度	得分					平均分
	1	2	3	4	5	
精神需求	0(0%)	2(2%)	29(29%)	42(42%)	27(27%)	3.94
物理需求	4(4%)	5(5%)	25(25%)	41(41%)	25(25%)	3.78
时间需求	0(0%)	9(9%)	40(40%)	32(32%)	19(19%)	3.61
表现	0(0%)	4(4%)	18(18%)	46(46%)	32(32%)	4.06
努力	0(0%)	0(0%)	28(28%)	48(48%)	24(24%)	3.96
挫折	2(2%)	16(16%)	37(37%)	24(24%)	21(21%)	3.46
合计	6(1%)	36(6%)	177(29.5%)	233(38.83%)	148(24.67%)	3.80

最后, 征集了用户对 BTVis 的意见和改进建议. 用户提出了一些宝贵意见, 包括需要加快交互的时间、给出交互加载进度、提供回退功能避免错误交互等. 未来, 将逐步添加这类增强系统可操作性和优化用户体验的功能.

4.4 定量实验

尽管定性实验和用户实验的结果已经对主题建模的结果质量进行了评估, 但是它们非常依赖人类的主观理解. 因此, 对系统性能的客观量化评估指标更加重要.

定量实验中, 除了上文中的 TED、豆瓣和大运河 3 个数据集外, 还采用 3 个真实数据集进行实验, 包括 20NewsGroup 新闻数据集、BBC_News 新闻数据集和 DBLP 论文文献数据集. 20NewsGroup 是一组在 NLP 领域中常用的新闻数据集, 共筛选 16 309 篇文档, 涉及 20 个不同的话题. BBC_News 是收集自 BBC 新闻网站的文本数据集, 共包含 2 225 篇新闻文章, 被分为 5 个不同的话题, 包括 business, entertainment, politics, sport 和 tech. BLP 是从 DBLP 官网内统计搜集得到的计算机领域内的论文文献数据集, 共包括 54 595 篇文档和 4 类标签. 本文中, 设置 20NewsGroup 数据集的主题数为 20, BBCNews 的主题数为 10, 大运河的主题数为 10, DBLP 的主题数为 30, TED 的主题数为 15, 豆瓣影评数据集的主题数为 40.

选择 9 种不同的方法进行实验, 包含了经典主题模型和最新的神经网络主题模型: (1) LDA[1]是一个经典的主题建模概率生成模型; (2) NMF[7]使用文档-主题和主题-单词 2 个小矩阵来分解文档-单词大矩阵; (3) ProdLDA[13]是使用 VAE 推断思想的自动变分编码推理主题模型; (4) ETM[52]是一种将传统主题模型与词嵌入相结合的文档生成模型; (5) ContextualNTM[57]是将上下文表示与神经主题模型相结合的情境化主题模型; (6) DeepNMF[58]是一种将多分支深度网络与 NMF 结合的无监督深度模型; (7) ContrastiveNTM[59]是对比解纠缠神经主题模型, 其将语义嵌入和对比学习融入变分自编码器中; (8) ECRTM[60]是嵌入聚类正则化主题模型, 用嵌入聚类正则化方法解决主题崩溃问题; (9) BERTopic[5]是一种采用嵌入、降维和聚类, 并使各操作部分模块化的神经网络主题模型.

评估模型时, 采用主题一致性、主题多样性和主题稳定性[19]指标. 其中, 主题一致性使用 NPMI 计算每个主题的前 10 个单词的关联程度; 主题多样性使用 Topic Diversity 指标衡量不同主题的重叠程度; 主题稳定性指模型在多次运行下主题间的相似程度.

在 6 个数据集上, 9 种方法的主题一致性、主题多样性和主题稳定性指标的评估结果如表 3 所示.

表 3　在 6 个数据集上 9 种方法的指标对比

方法	20NewsGroup			BBC_News			大运河		
	主题一致性	主题多样性	主题稳定性	主题一致性	主题多样性	主题稳定性	主题一致性	主题多样性	主题稳定性
LDA[1]	0.053	0.690	0.712	−0.021	0.380	0.746	−0.125	0.211	0.585
NMF[7]	0.083	0.735	0.791	−0.004	0.680	0.713	−0.067	0.187	0.561
ProdLDA[13]	0.020	0.875	0.645	−0.085	0.935	0.652	−0.049	0.310	0.716
ETM[52]	0.061	0.835	0.890	−0.013	0.840	0.842	−0.029	0.145	0.889
ContextualNTM[57]	0.048	0.880	0.694	0.080	0.800	0.652	−0.005	0.287	0.760
DeepNMF[58]	0.117	0.810	0.816	0.086	0.825	0.823	−0.001	0.311	0.882
ContrastiveNTM[59]	0.095	0.845	0.827	0.019	0.795	0.805	0.002	0.303	0.714
BERTopic[5]	0.120	0.850	0.859	0.047	0.850	0.840	0.010	0.336	0.921
ECRTM[60]	0.160	0.910	0.843	0.103	0.970	0.819	0.007	0.357	0.906
方法	DBLP			TED			豆瓣		
	主题一致性	主题多样性	主题稳定性	主题一致性	主题多样性	主题稳定性	主题一致性	主题多样性	主题稳定性
LDA[1]	−0.001	0.912	0.635	0.001	0.340	0.840	−0.102	0.620	0.624
NMF[7]	0.025	0.583	0.673	−0.002	0.580	0.743	−0.056	0.664	0.599
ProdLDA[13]	−0.090	0.775	0.692	−0.260	0.990	0.871	−0.148	0.872	0.675
ETM[52]	−0.005	0.550	0.812	−0.001	0.910	0.842	−0.004	0.843	0.892
ContextualNTM[57]	0.060	0.600	0.771	0.014	0.860	0.713	0.005	0.743	0.711
DeepNMF[58]	0.032	0.825	0.786	0.143	0.810	0.736	0.007	0.756	0.728
ContrastiveNTM[59]	0.021	0.795	0.742	0.118	0.845	0.732	0.004	0.782	0.763
BERTopic[5]	0.047	0.850	0.801	0.102	0.885	0.821	0.013	0.792	0.872
ECRTM[60]	0.075	0.894	0.801	−0.004	0.846	0.834	0.008	0.901	0.801

从表 3 可以看出: (1) 在不同数据集上, 各主题建模表现具有一定的差异性, 且没有能在不同数据集上同时保持最优性能的主题建模. (2) 主题一致性指标中, BERTopic[5]在短文本数据集中更具优势, 其每篇文档对应一个主题的策略, 使得在短文本中文档对应的主题更集中; DeepNMF[58]和 ContrastiveNTM[59]在长文本数据集中能获得更好的性能. (3) 主题多样性指标中, 由于 ECRTM[60]使每个单词尽可能地对应单个主题的建模策略, 因此可以保证生成的主题具有更高的多样性. (4) 主题稳定性指标中, BERTopic[5]和 ETM[52]的稳定性相对排名较高, 由于 BERTopic[5]每次运行表示文本的语义嵌入向量基本不变, 其聚类结果变化较小. (5) 在大运河数据集上, 由于该数据集的所有内容都与大运河高度相关, 因此所有模型的主题多样性远低于其他数据集. (6) 在不同数据集上, 神经网络模型的指标得分并不总是优于经典的概率模型. 例如, 在 TED 数据集上, 尽管 ProdLDA[13]获得了最高的多样性和稳定性, 但同时拥有最低的主题一致性, 表明该方法中的主题关键词具有独特性, 但是其关键词可能存在表述错误, 需要查看模型的主题细节才能进行进一步的分析.

本文通过不同指标下的定量实验对比评估发现, 模型并不能在所有数据集中一直保持优势, 这些模型在不同数据集中生成的基于深度神经网络主题建模可视分析的模型提升与对比的主题可能是低质量的, 妨碍用户理解, 在这种情况下, 通过定性分析来探索不同数据集上主题建模性能是重要的. 通过定性对比, 用户能够深入了解模型性能的局限性和潜在的错误, 揭示数据集内容的特定差异、模型假设的局限性、可能存在的噪声或干扰, 以及其他影响模型效果的因素.

5 结 语

本文开发了 BTVis, 用户可以通过可视化与 BERTopic[5]算法的结合逐步完善主题模型. 可视分析视图提供了各类分析, 包括中间结果概览、主题分析、文档概览和单词评估. 本文提出并实现了在 BERTopic[5]框架下的 3 种不同粒度的局部建模算法, 并能够根据用户需求循环地优化建模结果. 通过实验、案例分析和用户实验, 验证了通过该系统可以提高主题模型的质量与解释性, 辅助用户分析和挖掘主题信息.

但是, 本文工作也存在需要改进的地方: (1) 在实验过程中, 虽然语言大模型在大部分情况下可以可靠地概括主题, 但也存在可能生成错误标题、问答的反映时间较长等问题. 因此, 如何保证标题的可靠性和实时性是未来亟待解决的问题. (2) 与其他工作相似, 本文的比较方法仍然局限于在一个模型中分析不同的主题. 一位用户提出, 他想要比较不同主题模型在同一数据集上的建模结果, 并分析产生差异的原因. 事实上, 主题模型的结果会受到模型与数据集的双重影响, 在不同的数据集环境下, 使用不同的主题模型可能会得到更好的结果. (3) 目前很少有工作针对和比较不同主题模型间的结果, 且针对不同的无监督主题模型, 只能通过指标评价和人为评价 2 种评估方式. 从实际角度出发, 想要在实际应用中使用主题模型得到一个满意的结果需要花费大量时间分析和调参. 因此, 未来希望在本文工作的基础上比较不同的神经网络主题建模间的异同, 帮助用户在特定的环境中评估和选择更优的主题模型.

参考文献(References):

[1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993-1022

[2] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis[J]. Machine Learning, 2001, 42(1-2): 177-196

[3] Zhao H, Phung D Q, Huynh V, *et al*. Topic modelling meets deep neural networks: a survey[C] //Proceedings of the 30th International Joint Conference on Artificial Intelligence. Freiburg: ijcai.org, 2021: 4713-4720

[4] Sun F, Liu J, Wu J, *et al*. BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer[C] //Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2019: 1441-1450

[5] Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure[OL]. [2024-02-09]. https://arxiv.org/abs/2203.05794

[6] McInnes L, Healy J, Astels S. Hdbscan: hierarchical density based clustering[J]. Journal of Open Source Software, 2017, 2(11): Article No.205

[7] Lee D, Seung H S. Algorithms for non-negative matrix factorization[C] //Proceedings of the 13th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2000: 535-541

[8] Yan X H, Guo J F, Lan Y Y, *et al*. A biterm topic model for short texts[C] //Proceedings of the 22nd International Conference on World Wide Web. New York: ACM Press, 2013: 1445-1456

[9] Qiang J P, Qian Z Y, Li Y, *et al*. Short text topic modeling techniques, applications, and performance: a survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2022,

34(3): 1427-1445

[10] Diederik P K, Welling M. Auto-encoding variational Bayes [OL]. [2024-02-09]. https://arxiv.org/abs/1312.6114

[11] Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144

[12] Scarselli F, Gori M, Tsoi A C, *et al*. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2009, 20(1): 61-80

[13] Srivastava A, Sutton C. Autoencoding variational inference for topic models[OL]. [2024-02-09]. https://openreview.net/forum?id=BybtVK9lg

[14] Liu L Y, Huang H Y, Gao Y, *et al*. Neural variational correlated topic modeling[C] //Proceedings of the World Wide Web Conference. New York: ACM Press, 2019: 1142-1152

[15] Wang R, Zhou D Y, He Y L. ATM: adversarial-neural topic model[J]. Information Processing & Management, 2019, 56(6): Article No.102098

[16] Zhu Q L, Feng Z, Li X L. GraphBTM: graph enhanced autoencoded variational inference for biterm topic model[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 4663-4672

[17] Huang Jiajia, Li Pengwei, Peng Min, *et al*. Review of deep learning-based topic model[J]. Chinese Journal of Computers, 2020, 43(5): 827-855(in Chinese)
(黄佳佳, 李鹏伟, 彭敏, 等. 基于深度学习的主题模型研究[J]. 计算机学报, 2020, 43(5): 827-855)

[18] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 6000-6010

[19] Abdelrazek A, Eid Y, Gawish E, *et al*. Topic modeling algorithms and applications: a survey[J]. Information Systems, 2023, 112: Article No.102131

[20] Churchill R, Singh L. The evolution of topic modeling[J]. ACM Computing Surveys, 2022, 54(10s): Article No.215

[21] Thompson L, Mimno D. Topic modeling with contextualized word representation clusters[OL]. [2024-02-09]. https://arxiv.org/abs/2010.12626

[22] Angelov D. Top2Vec: distributed representations of topics[OL]. [2024-02-09]. https://arxiv.org/abs/2008.09470

[23] Le Q, Mikolov T. Distributed representations of sentences and documents[C] //Proceedings of the 31st International Conference on Machine Learning. New York: PMLR, 2014: 1188-1196

[24] Jiang Z H, Chen H D, Zhou R, *et al*. HealthPrism: a visual analytics system for exploring children's physical and mental health profiles with multimodal data[J]. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(1): 1205-1215

[25] Choo J, Lee C, Reddy C K, *et al*. UTOPIAN: user-driven topic modeling based on interactive nonnegative matrix factorization[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 1992-2001

[26] van der Maaten L, Hinton G. Visualizing data using *t*-SNE[J]. Journal of Machine Learning Research, 2008, 9(86): 2579-2605

[27] Ganesan A, Brantley K, Pan S M, *et al*. LDAexplore: visualizing topic models generated using latent dirichlet allocation[OL]. [2024-02-09]. https://arxiv.org/abs/1507.06593

[28] Tang Ying, Su Jianjing, Tong Ning. TMvis: a visual analysis system based on LDA topic modelling[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(10): 1728-1738(in Chinese)
(汤颖, 苏建明, 童宁. TMvis: 基于 LDA 的主题建模可视分析系统[J]. 计算机辅助设计与图形学学报, 2019, 31(10): 1728-1738)

[29] Zhu C, Cheng Y, Gan Z, *et al*. MaxVA: fast adaptation of step sizes by maximizing observed variance of gradients[C] //Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Research Track. Heidelberg: Springer, 2021: 628-643

[30] Dou W W, Liu S X. Topic- and time-oriented visual text analysis[J]. IEEE Computer Graphics and Applications, 2016, 36(4): 8-13

[31] Tang Siliang, Cheng Lu, Shao Jian, *et al*. A survey on news text visualization via probabilistic topic modeling[J]. Journal of Computer-Aided Design & Computer Graphics, 2015, 27(5): 771-782(in Chinese)
(汤斯亮, 程璐, 邵健, 等. 基于概率主题建模的新闻文本可视化综述[J]. 计算机辅助设计与图形学学报, 2015, 27(5): 771-782)

[32] Xiang Shouxing, Ouyang Fangxin, Liu Shixia. Concept-based visual analysis of dynamic textual data[J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(4): 531-541(in Chinese)
(向首兴, 欧阳方昕, 刘世霞. 概念层次的动态文本可视分析[J]. 计算机辅助设计与图形学学报, 2020, 32(4): 531-541)

[33] Zhang Yi, Shao Yudong, Zhang Jiawan. Visual analysis of topic transition among different sources of text corpora[J]. Journal of Computer-Aided Design & Computer Graphics, 2017, 29(12): 2265-2272(in Chinese)
(张怡, 邵裕东, 张加万. 多源媒体文本主题演变的可视分析[J]. 计算机辅助设计与图形学学报, 2017, 29(12): 2265-2272)

[34] Lyu Mengya, Wang Xiaolong, Li Kaixuan, *et al*. High-dimensional multi-attributeed data visual analysis system for negative comment guidance[J]. Journal of Computer-Aided Design & Computer Graphics, 2024, 36(6): 937-947(in Chinese)
(吕梦雅, 王晓龙, 李凯旋, 等. 负面评论引导的高维多元数据可视分析系统[J]. 计算机辅助设计与图形学学报, 2024, 36(6): 937-947)

[35] Poux-Médard G, Velcin J, Loudcher S. Information interaction profile of choice adoption [C] //Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Research Track. Heidelberg: Springer, 2021: 103-118

[36] Dou W W, Yu L, Wang X Y, *et al*. HierarchicalTopics: visually exploring large text collections using topic hierarchies[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2002-2011

[37] Hoque E, Carenini G. Interactive topic hierarchy revision for exploring a collection of online conversations[J]. Information Visualization, 2019, 18(3): 318-338

[38] Yang Y, Yao Q M, Qu H M. VISTopic: a visual analytics system for making sense of large document collections using hierarchical topic modeling[J]. Visual Informatics, 2017, 1(1):

40-47

[39] Yang W K, Wang X T, Lu J, *et al*. Interactive steering of hierarchical clustering[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(10): 3953-3967

[40] Kowsari K, Brown D E, Heidarysafa M, *et al*. HDLTex: hierarchical deep learning for text classification[C] //Proceedings of the 16th IEEE International Conference on Machine Learning and Applications. Los Alamitos: IEEE Computer Society Press, 2017: 364-371

[41] Zhang D C, Ying R, Lauw H W. Hyperbolic graph topic modeling network with continuously updated topic tree[C] //Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2023: 3206-3216

[42] El-Assady M, Sevastjanova R, Sperrle F, *et al*. Progressive learning of topic modeling parameters: a visual analytics framework[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 382-391

[43] El-Assady M, Sperrle F, Deussen O, *et al*. Visual analytics for topic model optimization based on user-steerable speculative execution[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 374-384

[44] El-Assady M, Kehlbeck R, Collins C, *et al*. Semantic concept spaces: guided topic model refinement using word-embedding projections[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 1001-1011

[45] Kim H, Drake B, Endert A, *et al*. ArchiText: interactive hierarchical topic modeling[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(9): 3644-3655

[46] Liu C Y, Li J, Shen T S. Refined semantic pattern interaction exploration for multi-attribute documents: a study on subspace topic modeling and visual analysis[J/OL]. Journal of Computer-Aided Design & Computer Graphics: 1-17[2024-02-09]. https://www.jcad.cn/article/doi/10.3724/SP.J.1089.2024-00108 (in Chinese)
(刘陈一, 李杰, 沈天舒. 面向多属性文档的精细化语义模式交互探索方法[J/OL]. 计算机辅助设计与图形学学报: 1-17[2024-02-09]. https://www.jcad.cn/article/doi/10.3724/SP.J.1089.2024-00108)

[47] Kuang D, Park H. Fast rank-2 nonnegative matrix factorization for hierarchical document clustering[C] //Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2013: 739-747

[48] Wu Yingcai, Cui Weiwei, Song Yangqiu, *et al*. A survey on topic-based visual text analytics[J]. Journal of Computer-Aided Design & Computer Graphics, 2012, 24(10): 1266-1272(in Chinese)
(巫英才, 崔为炜, 宋阳秋, 等. 基于主题的文本可视分析研究[J]. 计算机辅助设计与图形学学报, 2012, 24(10): 1266-1272)

[49] Liu S X, Wang X T, Collins C, *et al*. Bridging text visualization and mining: a task-driven survey[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(7): 2482-2504

[50] Ye Li, Wang Lei, Zhou Zhiguang, *et al*. A survey on the visual analytics of narrative structure abstraction[J]. Journal of Computer-Aided Design & Computer Graphics, 2024, 36(4): 503-515(in Chinese)
(叶力, 王磊, 周志光, 等. 基于叙事结构抽象的可视分析方法综述[J]. 计算机辅助设计与图形学学报, 2024, 36(4): 503-515)

[51] McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction[OL]. [2024-02-09]. https://arxiv.org/abs/1802.03426

[52] Dieng A B, Ruiz F J R, Blei D M. Topic modeling in embedding spaces[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 439-453

[53] Bouma G. Normalized (pointwise) mutual information in collocation extraction[C] //Proceedings of the German Society for Computational Linguistics and Language Technology. Tübingen: Gunter Narr Verlag, 2009: 31-40

[54] El-Assady M, Kehlbeck R, Metz Y, *et al*. Semantic color mapping: a pipeline for assigning meaningful colors to text[C]//Proceedings of the 4th Workshop on Visualization Guidelines in Research, Design, and Education. Los Alamitos: IEEE Computer Society Press, 2022: 16-22

[55] Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space[OL]. [2024-02-09]. https://arxiv.org/abs/1301.3781

[56] Hart S G. NASA-task load index (NASA-TLX); 20 years later[J]. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2006, 50(9): 904-908

[57] Bianchi F, Terragni S, Hovy D, *et al*. Cross-lingual contextualized topic models with zero-shot learning[C] //Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2021: 1676-1683

[58] Wang J Y, Zhang X L. Deep NMF topic modeling[J]. Neurocomputing, 2023, 515: 157-173

[59] Zhou X X, Bu J J, Zhou S, *et al*. Improving topic disentanglement via contrastive learning[J]. Information Processing & Management, 2023, 60(2): Article No.103164

[60] Wu X B, Dong X S, Nguyen T T, *et al*. Effective neural topic modeling with embedding clustering regularization[C] //Proceedings of the 40th International Conference on Machine Learning. New York: PMLR, 2023: 37335-37357