面向多属性文档的精细化语义模式交互探索方法

刘陈一, 李杰*, 沈天舒

(天津大学智能与计算学部 天津 300350)
(jie.li@tju.edu.cn)

摘 要: 现有的语义模式探索方式限制了对全局语义模式的理解, 少有研究关注保留上下文语义的精细化语义模式探索. 文中通过深度结合机器学习和可视化技术, 运用面向多属性文档的精细化语义模式可视分析方法, 使用户能够在灵活地分析不同属性下的精细化语义模式的同时感知全局语义模式. 首先引入表征重构网络, 得到包含多属性文档语义和属性信息的潜在向量, 使主题模型能够更好地识别子空间主题; 然后引入人在回路的语义模式可视分析方法, 开发了一套包括探索管理器、子空间投影仪和主题解释器的语义模式可视化系统, 支持用户选择表征重构的属性子空间, 交互探索语义模式, 并提供分析结果. 基于游戏评论数据集、美国新闻数据集和特朗普大选数据集, 采用主题多样性和主题一致性的指标与既有方法进行对比, 实验结果表明, 所提出的主题模型在主题建模上具有较好的泛用性和灵活性; 用户实验包括语义模式探索任务, 验证了所提方法和可视化系统在语义模式探索上的出色执行效率和有效性.

关键词: 语义模式; 多属性文档; 子空间主题建模; 可视分析; 交互探索
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024-00108

Refined Semantic Pattern Interaction Exploration for Multi-Attribute Documents: A Study on Subspace Topic Modeling and Visual Analysis

Liu Chenyi, Li Jie*, and Shen Tianshu

(*College of Intelligence and Computing, Tianjin University, Tianjin* 300350)

Abstract: Existing approaches to semantic pattern exploration limit the comprehension of global semantic patterns, with few studies focusing on the fine-grained semantic pattern exploration that preserves contextual semantics. In this paper, this paper combines machine learning and visualization techniques in depth, employing a fine-grained semantic pattern visual analysis method for multi-attribute documents. This allows users to flexibly analyze fine-grained semantic patterns under different attributes while perceiving global semantic patterns. Firstly, this paper introduces a representation reconstruction network to obtain latent vectors that contain semantic and attribute information of multi-attribute documents, enabling the topic model to better identify subspace topics. Then, this paper introduces a human-in-the-loop semantic pattern visual analysis method and develop a semantic pattern visualization system that includes an exploration manager, a subspace projector, and a topic interpreter. This system supports users in selecting attribute subspaces for representation reconstruction, interactively exploring semantic patterns, and providing analysis results. Based on game review datasets, US news datasets, and Trump election datasets, this paper compare our proposed topic model with existing methods using topic diversity and topic coherence metrics. Experimental results show that the proposed topic model has good generalizability and flexibility in topic modeling. User experiments including semantic pattern ex-

收稿日期: 2024-03-01; 修回日期: 2024-09-04. **基金项目**: 国家自然科学基金(62372321). **刘陈一**(2003—), 男, 在校学生, 主要研究方向为机器学习; **李杰**(1984—), 男, 博士, 副教授, 硕士生导师, CCF 会员, 论文通信作者, 主要研究方向为机器学习、大数据可视分析; **沈天舒**(1998—), 女, 硕士研究生, 主要研究方向为主题模型可视化.

ploration tasks verify the outstanding efficiency and effectiveness of the proposed method and visualization system in semantic pattern exploration.

Key words: semantic patterns; multi-attribute documents; subspace topic modeling; visual analysis; interactive exploration

1 相关工作

1.1 研究背景及意义

语义模式的探索是可视化中的经典研究问题. 给定一个带有多个属性的文档集合(多属性文档), 语义模式探索的目标是从中挖掘出主题, 并分析主题在某个属性(目标属性)上的流行度和内容是如何变化的[1]. 例如, 用户可以选择时间属性分析主题随时间的演化模式, 选择空间属性分析主题在不同区域上的空间分布.

大部分现有的语义探索工作都有 3 个主要步骤[2-3]: (1) 分析人员指定不同的查询条件以过滤出满足查询条件的文档, 可能的查询条件包括属性范围(带有特定属性值的文档)或关键词(带有特定关键词的文档); (2) 分析人员应用主题模型从过滤出的文档上提取主题; (3) 分析人员选择某个感兴趣的主题, 依据它们与文档之间的相关性计算该主题在某些文档属性上的分布, 该分布即为该主题在该属性上的语义模式.

上述语义模式探索方式容易忽视其中存在的一个不足: 分析人员每次在被过滤的文档上提取主题时, 无法得到不同批次提取出来的主题之间的关系, 限制了他们对全局语义模式的深入理解, 如图 1a 所示. 到目前为止, 鲜有研究关注对保留上下文语义的精细化语义模式探索的可能性, 以及这样做的潜在优势. 为了克服这种分步骤处理方式可能带来的问题, 本文提出标签生成和表征重构 2 个关键步骤, 将文档的属性过滤和语义模式挖掘融合到一个统一的任务中, 实现一种在单个上下文中得到多个子空间主题相关性的方法. 该方法能够在多属性文档中实现主题的精细化提取, 使分析人员既能关注于特定属性的语义模式, 又能在上下文范围内理解文档集合的整体性, 通过全局视角对文档进行语义分析, 如图 1b 所示.

进行语义模式探索时, 本文采取可视分析方法, 通过图形图像展示文本数据的结构和模式, 帮助分析人员快速地识别关键主题和语义模式; 同时, 利用缩放、平移和筛选等交互功能, 提供灵活地调整和分析数据的手段. 可视分析方法因其通用性, 可广泛适用于各类文本数据和语义模式探索任务, 为跨领域、多场景的多属性文档处理提供了统一且高效的方法论. 如果不使用可视化方法, 则将面临分析效率低下、人员培训成本高昂、结果理解困难以及方法通用性受限等挑战.

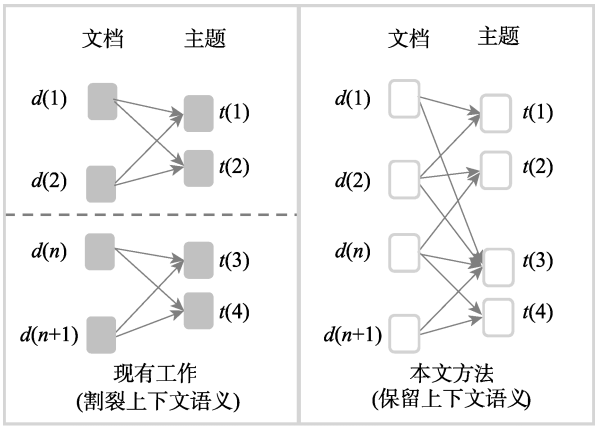


图 1 现有工作与本文方法主题对比

1.2 国内外研究现状及分析

1.2.1 主题模型

当前, 支持多属性文档子空间内的主题提取和语义模式挖掘的主题模型相关工作划分为 4 大类别: 传统主题模型、层次结构主题模型、有监督的主题模型和基于嵌入的主题模型.

主题建模是一种文档软聚类, 可以得到文档对主题的概率分布[4], 利用术语频率-逆文档频率(term frequency-inverse document frequency, TF-IDF)[5]评估文档的主题. 传统的主题模型基于词袋表示, 有 2 类: (1) 概率模型. 如潜在狄利克雷分配(latent Dirichlet allocation, LDA)主题模型[6], 使用狄利克雷先验建模主题-关键词和文档-主题分布; (2) 非概率模型. 其中最突出的是潜在语义索引(latent semantic indexing, LSI)[7]非负矩阵分解(non-negative matrix factorization, NMF)[8]. 然而, 大部分现有的传统主题模型都无法关联各个子空间文档, 缺少全局语义的相关性.

层次结构主题模型是一种用于发现主题之间关

联和层次结构的方法. Blei 等[9]采用树状结构表示主题的组织和构建；层次狄利克雷过程(hierarchical Dirichlet process, HDP)[10]能够自动确定聚类(主题)数量. 但是，传统的层次主题模型通常不能直接捕获与特定子空间相关的主题.

有监督的主题模型的特点是使用带有标签的文档作为输入，引导主题模型更好地捕捉文档的主题信息. 经典的有监督主题模型有 DiscLDA[11]和 Supervised LDA[12]等，但这些方法的标签一般无法得到标签文档和全部文档的语义关系.

随着人工智能技术的普及，出现了很多基于嵌入的主题模型[13-14]. 在此基础上，Angelov[15]提出 Top2Vec，将文档和词联合嵌入，并寻找文档的聚类质心代表主题嵌入；Grootendorst[16]提出了 BERtopic 方法，Egger 等的实验[17]证明了该方法在简短和非结构文本上的表现优于 Top2Vec.

与现有的主题模型工作相比，本文方法将标签解释为文档在全部文档中所属的子空间，以保留全局语义，进而提高主题的语义多样性；同时，强调在主题建模中考虑用户需求的可能性，提出子空间主题建模方法，便于构建灵活的主题模型并控制主题结果.

1.2.2　语义模式探索

语义模式挖掘是近年来的一个研究热点，可以帮助人们深入理解和挖掘文本数据中的潜在语义结构和模式. 常见的语义模式挖掘方法包括但不限于频繁项集挖掘、关联规则挖掘和序列模式挖掘等[18-20].

第 1 种基于词频的模式挖掘方法通过关键词过滤文本数据. Cao 等[21]的方法允许用户指定关键词，以识别主题在不同属性上的语义模式; Hu 等[22]使用频繁的顺序模式挖掘算法自动识别文本数据中的关键词: 然而，这些方法在解释文档的内容方面语义范围有限.

主题建模在语义模式探索中得到了广泛的应用. Alexander 等[23]采用邻接矩阵描述输入文档和主题之间的相关性. 在可视化方面，平行坐标图和语义-时空立方体[3,24]被应用于呈现文档数据在不同属性条件下的分布: 然而，这些方法无法提供有关全局语义模式的全面信息.

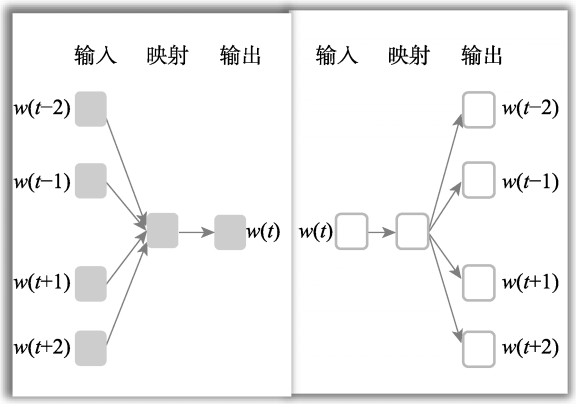
另一种常见的方法是基于机器学习的模型，它们专门设计了特定的分析框架. RoseRiver[25]可视化不同粒度级别主题的演化；TopicPanorama[26]揭示不同语料库之间主题演化的层次结构. 需要注意的是，上述方法通常是针对特定场景或任务设计的，在处理多属性文档子空间时，需要考虑如何整合它们的框架以实现语义模式探索.

一些工作通过文档嵌入的方式，使用视觉浏览和分析空间表示实现语义模式探索. Kim 等[27]的工作支持用户在初始文档嵌入的基础上，通过镜头界面进行动态的语义模式探索. 本文方法使用多属性文档的不同属性表示不同的分析任务，以重构文档嵌入，力求把语义模式探索任务泛化，实现一种通用化的语义模式探索方法.

1.3　基于嵌入的主题模型

1.3.1　词嵌入和文档嵌入的基本概念

嵌入概念最早出现于 Bengio 等[28]提出的神经网络语言模型(neural network language model, NNLM)，其将单词表示为较低维度向量以表达它们的语义特征. Word2Vec 模型[29]使用周围的 n 个单词表示目标单词，其单词向量模型包括连续词袋(continuous bag of words, CBOW)模型和连续跳字模型(continuous skip-gram model, Skip-gram)这 2 种方法，如图 2 所示. Le 等[30]提出 Doc2Vec，其扩展了 Word2Vec 模型，可以用于处理整个文档的语义信息[31].



a. CBOW 模型　　b. Skip-gram

图 2　Word2Vec 模型结构

Devlin 等[32]提出了图 3 所示 BERT 模型，其采用 Transformer[33]编码器作为主要框架，能够捕捉句子中的双向关系. Reimers 等[34]提出 Sentence-BERT，使用一种 Siamese 网络架构[35]提高句子级别的语义嵌入质量. 近年来，随着深度学习的迅猛发展，越来越多的语义表征技术被应用于改进主题模型，以提高主题提取的质量.

1.3.2　经典的嵌入式主题模型

嵌入式主题模型通常更适用于结合可视化，本文使用嵌入式的主题模型帮助用户理解文本数据的主题结构，提供更直观的主题信息.

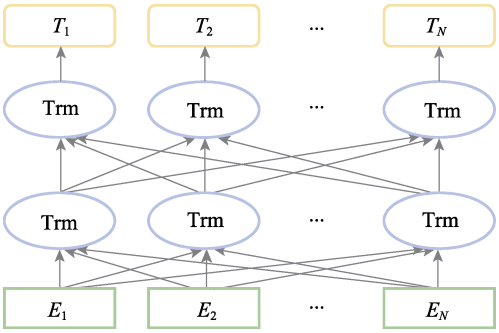
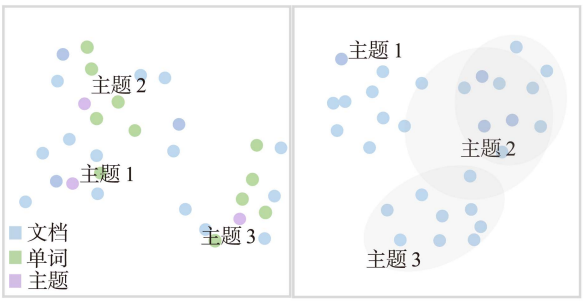


图 3 BERT 模型结构

(1) Top2Vec. 该模型假设文档和词在同一个向量空间，如图 4a 所示. 首先将文本数据嵌入到高维空间中，使用 Doc2Vec 或其他嵌入模型捕捉文档的语义信息；然后使用聚类算法对文档向量进行聚类，产生一组由相关文档组成的主题簇. 由主题向量相邻的词向量表示主题关键词，同时使用均匀流形近似与投影(uniform manifold approximation and projection, UMAP)[36]等降维技术实现主题的可视化.

(2) BERTopic. 该模型通过对 Sentence-BERT 模型进行预训练获得高质量的文档嵌入，并且只训练文档的嵌入向量，如图 4b 所示. 与 Top2Vec 可以直接得到主题词不同，BERTopic 需要根据聚类结果使用 TF-IDF 生成主题词，解决了 Top2Vec 由 Doc2Vec 带来的语义歧义性的问题.



a. Top2Vec[15] b. BERTopic[16]

图 4 2D 嵌入空间示例

1.4 人在回路的主题模型

人在回路(human-in-the-loop, HITL)的主题模型旨在实现交互式修改主题模型的建模结果[37-40]，需要借助可视化和交互界面使用户理解主题模型的结果、改进模型，如图 5 所示. 值得注意的是，El-Assady 等[41]和 Dou 等[42]的方法均允许用户参与到建模过程中，增强了主题模型的灵活性和实用性.

HITL 的主题模型将人的主观性和领域知识与主题建模相结合，用户积极参与到主题模型的构建和解释中，包括定义主题、选择特征词汇、解释主题的语义和调整模型参数等[43-45]. 然而，已有的工作方法侧重于细化和调优主题模型，而不是为了获得与子空间相关的主题. 本文方法具有独特性，因为它旨在引导用户以交互的方式获取与特定子空间相关的主题，从而更好地满足特定领域对主题模型精细化探索的需求.

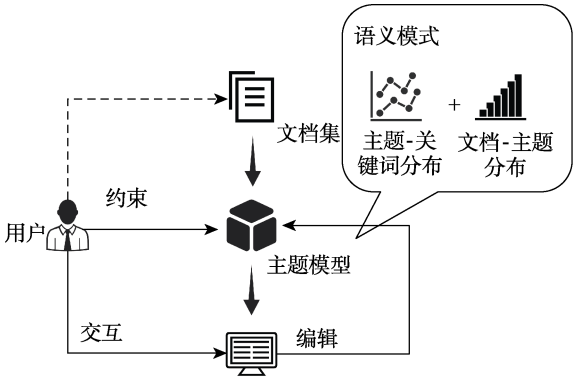


图 5 HITL 的主题模型提取主题的示意图

1.5 主题模型可视化

本文使用 3 种主题模型可视化方法:

(1) 属性信息的树形图. 树形图由 Johnson 等[46]提出，是一种流行的利用包含关系表达层次化数据的可视化方法[47]. 隐式树形图在解释树结构方面给用户带来了很高的负担[48]. 因此，本文使用显式树形图为用户提供对子空间标签进行灵活编辑的可能性.

(2) 文档信息的散点图. 本文运用文档嵌入散点图表示语义空间[49-52]，散点图的点的位置同时映射文档的语义和属性特征的分布，点的颜色对应树形图标签分布.

(3) 主题信息的柱状图. 在主题模型可视化中多用作展示文档、主题和关键词分布[53-55]. 本文中，柱状图的条形长度映射主题中不同标签的数量，用于表示主题随时间、地点或其他属性的变化趋势.

具有多个协调可视化方法的主题模型可视化系统被认为对探索复杂的数据集是有效的[56]. 本文使用树形图、散点图和柱状图，并联合使用多种可视化方法构建主题模型可视化工具，设计简洁的用户交互方法增强主题模型.

2 本文方法概述

本文提出一种向多属性文档的精细化语义模式可视分析方法，使用户能够在保留上下文语义的同时，灵活地分析不同属性下的精细化语义模式，感知全局语义模式. 首先拓展 BERTopic 提出

一种子空间主题建模方法，实现子空间语义探索任务的神经网络模型；其次将嵌入式主题模型集成到一个可视化系统中，使得分析人员能够进行复杂的探索任务，集成上文的子空间主题建模方法，实现 HITL 的主题模式挖掘工作流；最后通过 3 组关于使用可视化系统进行语义模式探索的案例、一项关于特定语义探索任务的用户实验和 2 组量化实验，证明了本文方法和可视化系统的有效性和可用性.

3　子空间主题建模

3.1　子空间的定义

设 $D(A_1,A_2,\cdots,A_n)$ 是一个包含 n 个属性文档集合，其中，A_i 表示一个值域为 B_i 的属性．D 中的每个元素 d 表示一个文本，并在每个属性上存在一个值 $d(A_i)$ 被包含于 B_i．图 6 所示为一个多属性文档的例子.

一个多属性文档集合包含多个子空间，每个子空间 S 对应一组过滤器，即 $r(A_1),r(A_2)\cdots r(A_n)$，其中，$r(A_i)$ 表示一个过滤器，用于指定在属性 A_i 上的一个值范围，即 B_i．则 S 包含这些过滤器选出的文档集合，其定义为

$$S=\{d\,|\,d(A_i)\in r(A_i),1\leqslant i\leqslant n\} \tag{1}$$

本文将覆盖整个域的过滤器(即 $r(A_i)$)隐藏，并将未隐藏的过滤器的属性作为子空间属性；全部未隐藏过滤器的数量表示子空间的维度；由未隐藏过滤器选取的属性值组合重构形成子空间．例如，图 6 中，年份属性设置为 2016 的 1D 子空间(如蓝色虚线框所示)和年份为 2016 且月份为 2 的 2D 子空间(如绿色虚线框所示)，因此，在多文档属性文档的某个子空间内提取主题的方法获得的主题的语义更具体．可以发现，从 2D 子空间得到的主题分别与更具体的事件相关.

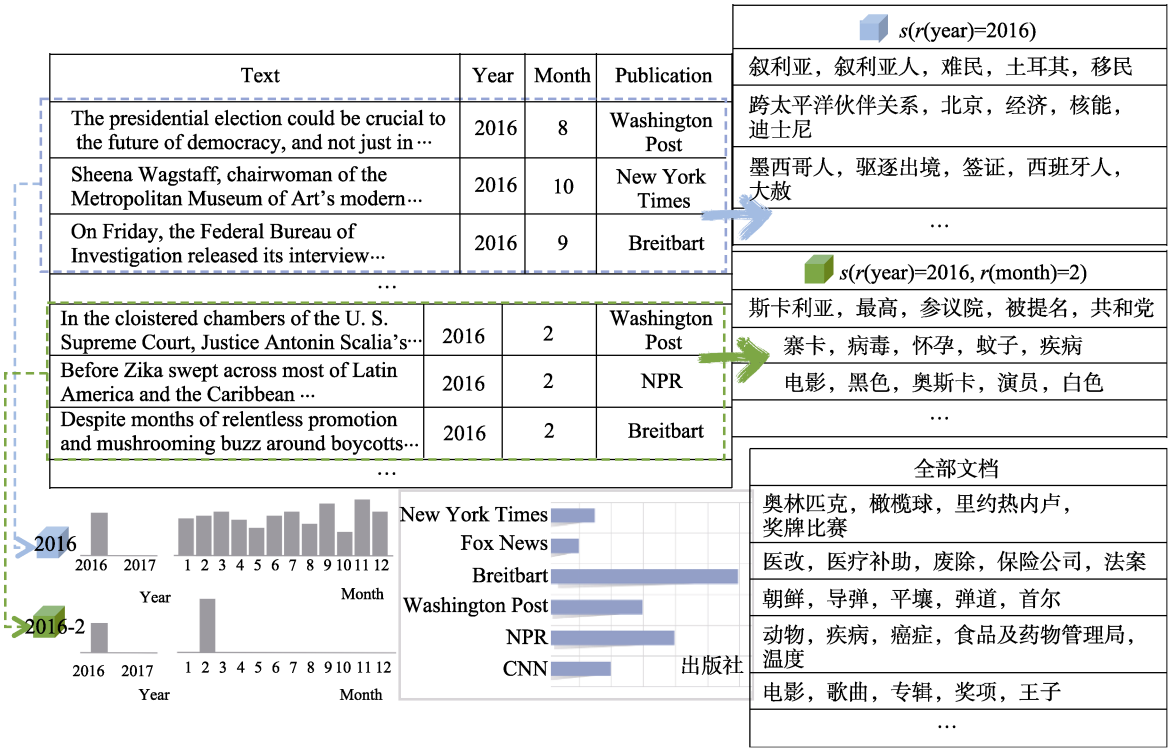


图 6　全局主题与子空间主题对比

3.2　设计思想

本文基于 BERTopic 提出一种方法，用于从全部文档中提取与特定子空间相关的主题，其包含 4 个步骤，如图 7 所示.

Step1. 使用 SBERT(sentence-bidirectional encoder representations from transformers, SBERT)模型为每个文档生成一个 384D 的嵌入；为了对每个文档生成更加紧凑的嵌入，使用 UMAP 将 384D 的文档嵌入映射到一个低维的潜空间；同时，根据用户的子空间属性选择为文档分配标签．该过程如图 7 中的 a 部分所示.

Step2. 重构文档表征，将属于相同子空间的文档的嵌入在潜空间中拉近，如图 7 中的 b 部分所示.

Step3. 用户对重构后的文档嵌入进行手动聚类，每个聚簇对应一个主题，根据主题的 TF-IDF 分数表示相关性，如图 7 中的 c 部分所示.

Step4. 利用可视化和交互探索不同主题在全部文本空间上的语义模式，如图 7 中的 d 部分所示. 图 7 中，标记了“*”的标签生成和表征重构是对 BERTopic 的扩展.

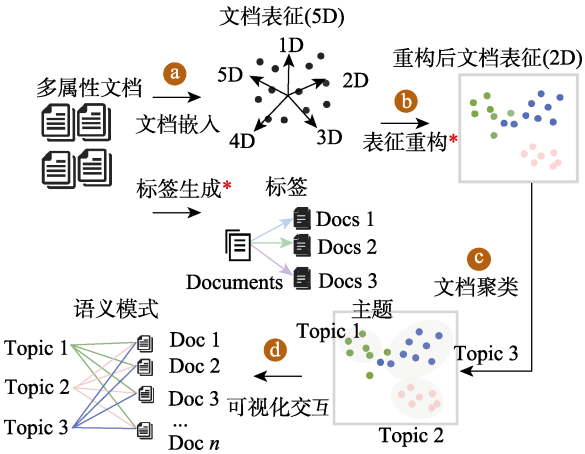


图 7 子空间主题建模方法

原始的 BERTopic 方法只包含文档嵌入、降维和文档聚类 3 个步骤，受限于其文档嵌入和聚类方法，当数据集包含复杂多样的语义信息时，可能无法充分地覆盖所有重要的主题，影响主题挖掘的全面性和深度. 为了在全局语义模式下实现语义信息的精细化探索，本文提出标签生成和表征重构 2 个关键步骤，通过拉近潜空间中相似子空间的文档嵌入，使得具有相似子空间属性的文档更容易形成紧密的聚簇，提高主题挖掘的准确性，并增强了主题的多样性和可解释性. 此外，通过表征重构，用户能够在整个文档集合上执行聚类，该方法覆盖数据集中的多样化信息，为用户提供了更全面的视角和更深入的数据洞察.

3.3 标签生成

本文采用属性值域的离散化方法实现用户定义子空间属性. 对于分类属性，直接使用其原始属性值；而对于数值属性，运用分箱技术将其连续的属性值范围划分为离散的区间.

本文根据每个文档的子空间属性的组合为其分配一个标签，如图 8 所示，每个标签对应一个包含具有相同标签的文档构成的属性子空间. 因此，每个文档都具有一个由 SBERT 生成的文档嵌入，以及一个指示其所属子空间的标签. 这种精细的标签设计有助于更准确地定义文档所属的属性子空间，优化主题建模的质量.

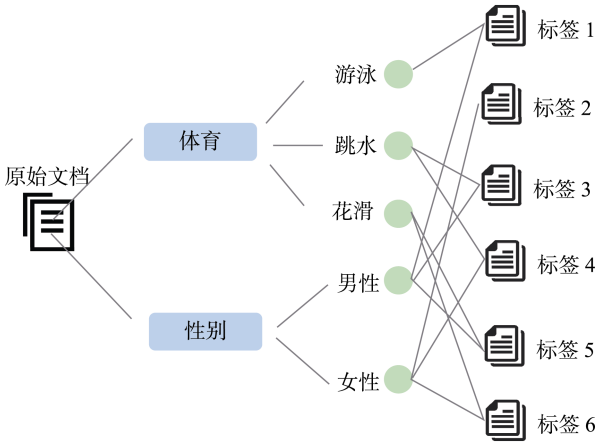


图 8 标签示意图

分别从不同属性值的文档集合中提取主题可以获得特定属性值下的语义模式，深入理解这一子空间内文档的特征. 与在不同子空间中分别获得互相独立的主题不同，本文方法利用标签对文档表征进行重构，实现了对整体文档集合的有效划分，并能够快速地提取具有全局性的主题.

3.4 表征重构

使用本文标签生成方法为每个文档设置标签，如图 9a 所示. 每个文档 x 同时具有一个文档嵌入 $\boldsymbol{h}$ 和一个指示其所属子空间的标签 $\boldsymbol{y}$；其中，$\boldsymbol{h}$ 由 SBERT 生成的原始文档嵌入降维得到.

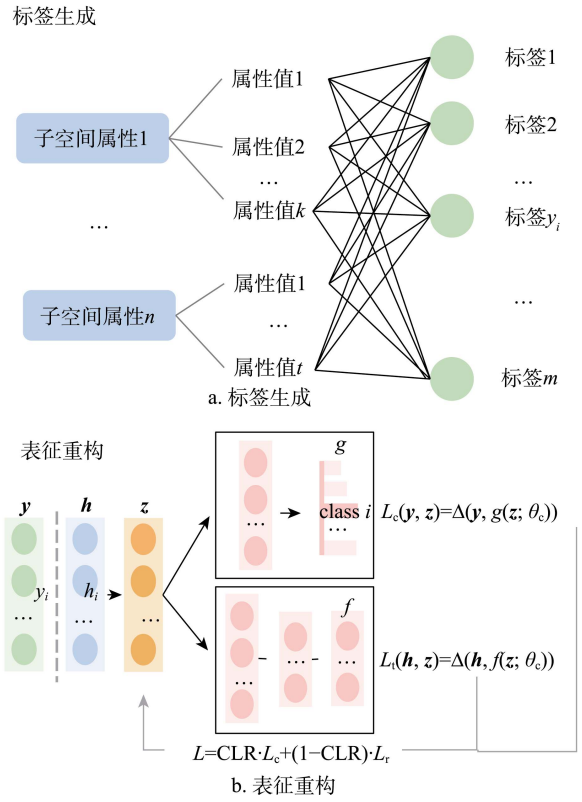


图 9 表征网络的结构

本文训练完成的表征重构网络利用 $\boldsymbol{y}$ 和 $\boldsymbol{h}$ 为 x 学习一个重构表征 $\boldsymbol{z}$，用于综合反映文档的主题和类别信息，图 9b 所示为表征重构网络的结构. 受到多视图嵌入[57]的启发，该网络使用一个双子网的结构，2 个子网共享重构嵌入 $\boldsymbol{z}$ 作为输入并分别预测文档类别 $\boldsymbol{y}'$ 和重构嵌入 $\boldsymbol{h}'$. 本文方法使用 $\boldsymbol{h}$ 初始化 $\boldsymbol{z}$，并在训练期间迭代更新 $\boldsymbol{z}$.

本文方法通过联合使用全连接网络中的重构损失和分类器中的分类损失迭代训练文档的重构的嵌入.

(1) 重构损失 L_{r}. 用于减少原始嵌入 $\boldsymbol{h}$ 与重构后的嵌入 $\boldsymbol{h}'$ 之间的差异，公式为

$$L_{\mathrm{r}}\left(\boldsymbol{h},\boldsymbol{z}\right)=\Delta\left(\boldsymbol{h},f\left(\boldsymbol{z};\theta_{\mathrm{r}}\right)\right) \tag{2}$$

其中，$f\left(\boldsymbol{z};\theta_{\mathrm{r}}\right)$ 表示预测 $\boldsymbol{h}'$ 的函数；θ_{r} 表示函数参数.

(2) 分类损失 L_{c}. 用于减少原始标签 $\boldsymbol{y}$ 与重构后的标签 $\boldsymbol{y}'$ 之间的差异，公式为

$$L_{\mathrm{c}}\left(\boldsymbol{y},\boldsymbol{z}\right)=\Delta\left(\boldsymbol{y},g\left(\boldsymbol{z};\theta_{\mathrm{c}}\right)\right) \tag{3}$$

其中，$g\left(\boldsymbol{z};\theta_{\mathrm{c}}\right)$ 表示预测 $\boldsymbol{y}'$ 的函数；θ_{c} 表示函数参数.

(3) 本文联合使用 L_{c} 和 L_{r} 作为嵌入网络的损失[58]，公式为

$$L=\mathrm{CLR}\cdot L_{\mathrm{c}}+\left(1-\mathrm{CLR}\right)\cdot L_{\mathrm{r}} \tag{4}$$

其中，分类损失比(classification loss ratio, CLR)是调整 L_{c} 和 L_{r} 的百分比的参数，$\mathrm{CLR}\in[0,1]$. 通过联合使用 L_{c} 和 L_{r}，文档表征之间距离同时反映了它们在语义和类别上的相似性.

本文采用一个只包含单个隐含层的全连接神经网络作为重构嵌入 $\boldsymbol{h}$ 的子网 f，并将预测文档类别 $\boldsymbol{y}$ 的子网 g 作为分类器. 使用均方误差(mean squared error, MSE)[57]作为 L_{r}，并使用 Li 等[58]提出的分类损失作为 L_{c}，它比现有的分类损失更容易在投影中产生聚簇等可视化结构.

本文的表征网络训练流程如图 9b 所示，该模型接受分析人员提供的分类标签 $\boldsymbol{y}$ 和文档嵌入 $\boldsymbol{h}$ 作为输入. 在初始训练阶段，首先随机初始化其模型参数以及表征向量 $\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_k\}$；然后利用随机生成的表征向量 $\boldsymbol{h}$ 和网络的参数 θ 构建表征网络模型. 在整个训练过程中，该网络使用式(2)中的重构损失 L_{r} 更新网络参数 θ；同时使用式(4)中的重构损失和分类损失的加权和 L 更新表征向量 $\boldsymbol{h}$，训练持续进行至总损失 L 不再显著减少为止. 一旦训练完成，该网络就能够直接获取表征向量 $\boldsymbol{h}$. 在本文具体实施中，设定文档嵌入 $\boldsymbol{h}$ 的维度为 5.

表征重构对主题的影响可以分成 4 种情况，如图 10 所示. 可以看出，图 10a 中，重构之后的表征依然相似；图 10b 中，重构之后的表征会形成几个具有相同标签的聚簇；图 10c 中，重构之后的嵌入变得相似，并有可能形成聚簇，而是否形成聚簇将由文档的语义相似程度决定；图 10d 中，不会将这些文档识别为聚簇(主题).

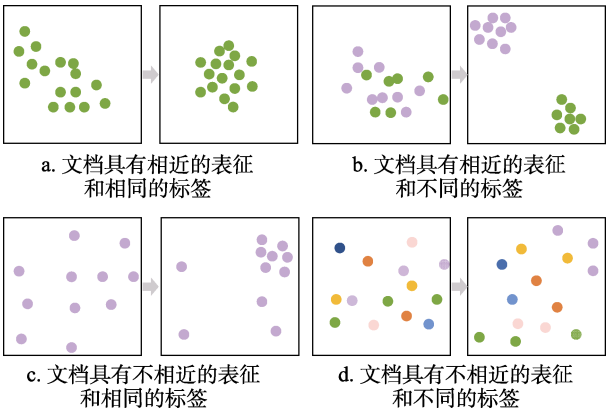


图 10　4 种重构情况

4　HITL 的语义模式可视分析系统

4.1　设计需求

本文深入研究了多属性文档语义模式探索方法的现有文献和工具. Hu 等[37]指出，现有主题模型的目标函数通常缺乏足够的上下文信息来完全理解数据的复杂性；El-Assady 等[38]和 Park 等[39]的工作均提出了对用户驱动的主题建模可视化系统的需求. 在主题建模工作中，需要层次化主题结构、精确的主题空间可视化和清晰的用户交互反馈[38-40]. 据此，本文分析并总结出了一系列关键需求：(1) 特定子空间相关联主题的获取；(2) 保留上下文语义的多属性文档语义模式探索；(3) 清晰有效的交互反馈.

为了确保设计需求与实际应用场景相契合，本文与多位对多维数据可视化分析工具使用具有丰富经验的用户和分析师进行了深入的访谈和讨论，以进一步验证和完善可视系统设计. 为了支持用户进行快捷、强大的多属性文档的语义模式探索，可视系统包含以下功能：(1) 支持以层次的结构组织探索任务；(2) 利用属性管理实时对属性进行筛选；(3) 清晰、直观地展现多属性文档语义和文档内容；(4) 获取感兴趣子空间的主题-属性分布，以一致的形式支持探索任务；(5) 提供简洁的

可视化界面和丰富的交互功能.

4.2 系统概述

图 11 所示为本文系统的流程, 其可实现基于重构的语义模式探索. (1) 分析人员根据用户指定属性将多属性文档划分为子空间; (2) 查看子空间重构的文档嵌入上的主题分布; (3) 对分析主题的语义模式进行解释. 该系统整合了探索管理器(exploration manager, EM), 子空间投影仪(subspace projector, SP)和主题解释器(topic explainer, TE)这 3 个核心组件, 可视化界面如图 12 所示.

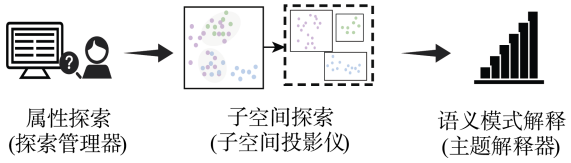


图 11 本文系统的步骤及组件

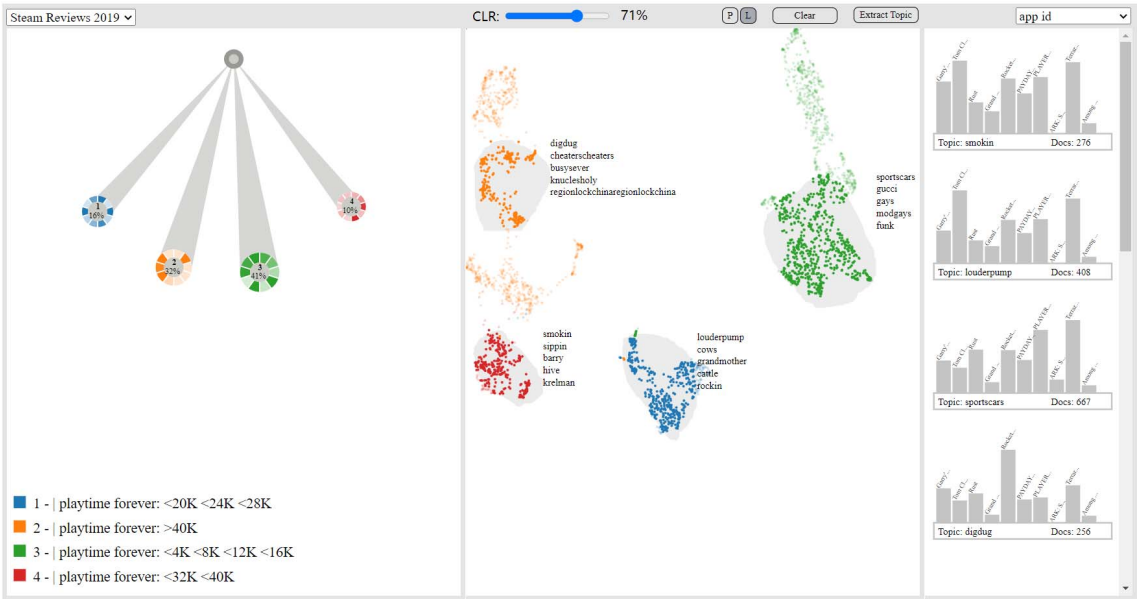


图 12 本文系统可视化界面

如图 12 最左栏所示, EM 的主要任务是生成子空间并将文档属性进行标签化, 用户可以通过 EM 在指定属性上组合不同的标签, 以进行语义模式的分析. 如图 12 中间栏所示, SP 负责创建子空间重构后的文档嵌入的潜在空间, 通过交互反映用户的需求, 帮助理解文档和属性的上下文语义并挖掘文档潜在的聚簇, 即主题. 如图 12 右侧所示, TE 与子空间探索操作相结合, 用一系列属性的量化指标协助分析人员理解与所选主题相关的语义模式, TE 展示了一组主题的属性分布.

4.3 系统界面与功能

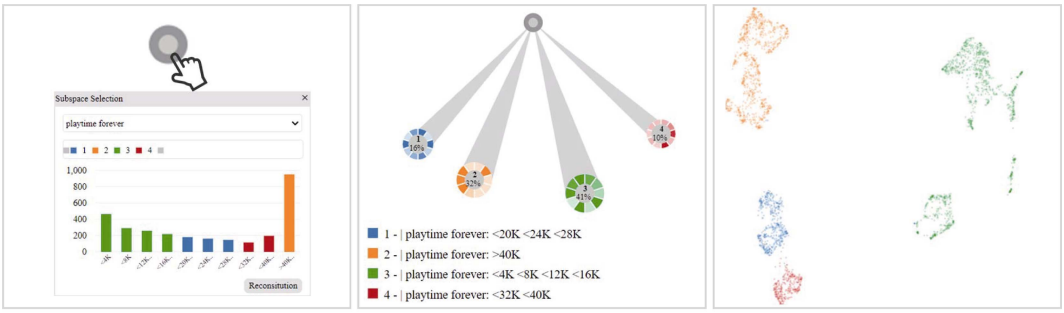
4.3.1 EM

EM 界面如图 12 所示, 其采用一种渐进的树形数据分区方法分析语义模式. 该方法通过显式的树形图实现, 其中, 每个树节点表示不同的属性, 而叶子节点则表示标签, 不同层之间的叶子节点互不包含相同的文档. 此外, 树形图下方的条形图用于展示和操作数据聚类结果. 下面详细描述 EM 的视觉设计及其操作方法.

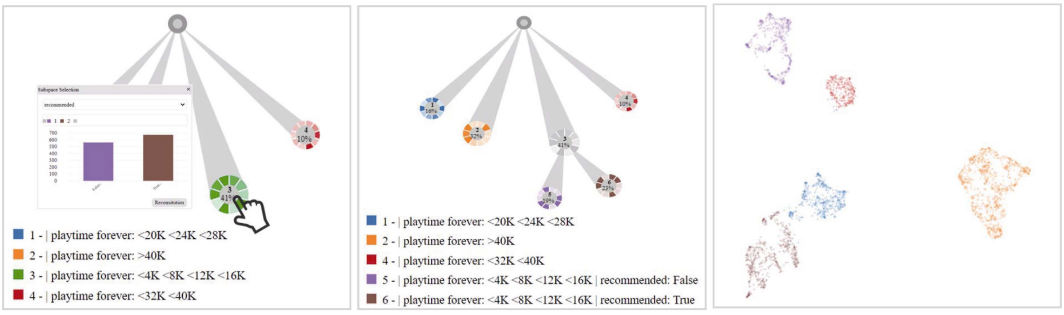
(1) 属性子空间创建. 本文采用一种树状结构的比喻作为 EM 的设计基础. 初始时的灰色的根节点代表整个文档数据集, 在此根节点触发的弹出窗口(如图 13a 所示), 分析人员可以选择并定义属性, 通过条形图呈现创建的属性分组操作结果.

本文为每条记录(数据)构建向量 $\boldsymbol{V}(\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_n)$, 以便进行聚类分组, 将向量均值作为该分组的向量表示, 最后采用 k 均值聚类方法[59]对这些向量进行聚类, 生成新的分组. 分析人员能够调整条形的分配, 如图 14 所示, 其中, 每个条形图簇对应一个属性区间, 相同区间内的条形颜色相同. 在条形图簇中, 分析人员能够通过交互组件添加分组, 并将条形图分配给指定分组; 此外, 还可以过滤掉不需要的分组.

EM 将每个属性区间编码为树的一个叶子节点, 如图 13 所示. 该界面中, 节点表现为一个饼状图, 饼状图的每个切片对应文档嵌入的一个维度, 而切片的颜色则编码该属性区间在这个维度上所



a. 属性子空间创建



b. 多属性子空间创建

图 13　EM 功能操作



图 14　分组控件的使用情况

有表征的平均值，颜色越深，表示在该维度上的平均值越高．饼的大小对应区间内的文档数量，饼的中心呈现类别索引和文档数量占比．底部的图例展示了每个扇形的颜色以及属性区间．

(2) 多属性子空间创建．如图 13b 所示，图 13a 中获得的饼图可作为一个全新的根节点，并在其上创建新的属性标签．这种交互方式允许分析人员渐进地探索和管理不同子空间，每个子空间都被分配一个独特的颜色；带有颜色的叶节点表示有效的属性子空间，如图 13b 中包含的 5 个有效子空间．这种树形结构的视觉表示方式有助于用户清晰地理解子空间之间的层次关系和结构．

(3) 数据过滤．当分析人员使用弹出窗口创建类时，可以通过滤出条形图将与所创建的类不相关的文档从数据集中排除．这种过滤机制允许分析人员更精确地定义和控制所关注的数据集．

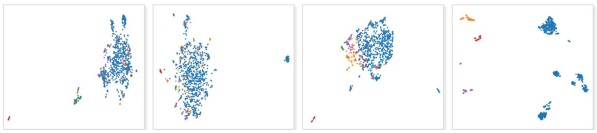
4.3.2　SP

SP 带有一组交互功能，利于语义内容的观察和主题的选择，其后台运行的嵌入网络接收到来自 EM 的标签后会生成一个重构网络．SP 使用 UMAP 技术降维创建投影，同时提供了基本的视图操作功能，如缩放、平移和套索选择．在生成的投影中，每个点表示一个文档，而点的颜色与 EM 中属性子空间的颜色相对应，如图 12 所示．

如图 12 所示，通过 SP 顶部的一个滑块可以随时调整参数 CLR．当 CLR 设置为 0 时，这种设置呈现数据集最初的语义模式；将 CLR 设置为 100% 时，则呈现不同属性子空间的语义模式；通常，CLR 应该位于这 2 个极端值之间，以充分地考虑属性信息对数据分布的影响．

图 15 所示为不同 CLR 值对数据分布的影响，其中，文档被分为 5 组，并使用不同的颜色表示．可以看出，随着 CLR 值的增加，文档聚类逐渐变得清晰，表明在语义上相邻的聚类之间具有更高的相似性．

在文档嵌入上能够查看文档内容，分析人员可以根据投影的分布情况，自由选择文档簇以及选择集群或异常值等分布进行进一步的分析．能



a. CLR=0 b. CLR=50% c. CLR=80% d. CLR=100%

图 15 不同 CLR 对数据分布的影响

够提取感兴趣的文档簇的主题，通过后端进行主题建模，该主题的前 5 个关键词将显示在主题投影的右侧，如图 12 所示.

4.3.3 TE

在 SP 中，分析人员可以使用套索工具选择主题(如图 12 中的灰色阴影区域)，然后利用 TE 进行语义模式的探索和比较. TE 允许分析人员选择一个特定的属性，查看所有主题在该属性上的分布情况，呈现为柱状图的形式，如图 12 所示.

流行度分数是主题在某个特定属性(如空间地点、时间等)上的流行程度或受欢迎程度的量化指标. 本文引入流行度分数的概念，探索主题的语义模式在不同属性上的分布情况. 计算每个主题与在特定属性上具有相同属性值的文档之间的相关性之和，将其作为该主题在该属性值上的流行度分数. 这种方法能够揭示主题在不同属性上的流行程度和变化趋势，为分析人员提供更全面的主题语义模式理解.

图 16 所示为语义模式的计算方法. 令 t 表示一个主题，w_{mk} 表示第 m 个文档与第 k 个主题之间的相关性. 取得全部在属性 A 上值为 a_n 的文档并构成了子空间(标记为浅蓝底色)，计算这些文档与主题的相关性之和作为 t 在 a_n 上的流行度分数(popularity score, PS). 本文使用相同的方法，得到 t 在 A 的每个属性值上的流行度分数.

TE 展示了主题在整个数据集和属性子空间的分布情况. 通过 TE，用户可以深入挖掘主题在不

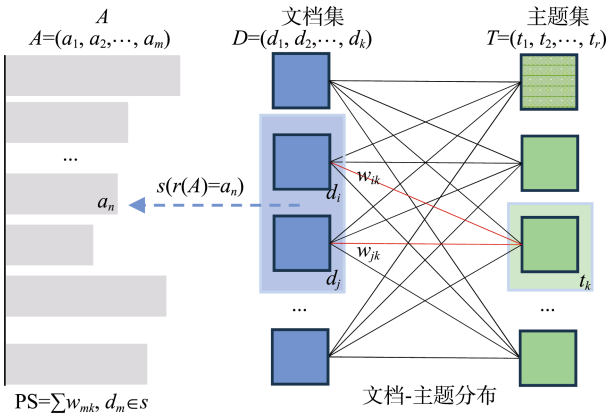


图 16 语义模式探索原理

同属性上的特征，发现不同属性之间的关系，以及主题随时间、地点或其他属性的变化趋势.

4.4 系统实现

本文系统采用 HTML/CSS/JavaScript 技术栈开发前端界面，后端服务程序通过 Python 编程语言编写，前后端之间的通信过程由轻量级的 Flask 框架实现；可视化视图使用 JavaScript 库 D3.js[60]创建和呈现. 本文使用 TensorFlow 框架[61]表征网络训练. 通过客户端-服务器架构、前后端技术的结合，以及数据可视化和深度学习框架的运用，该系统能够在多个浏览器中高效运行，并为用户提供数据分析和可视化功能.

5 实验评估

5.1 案例及结果分析

本节通过 3 个案例，展示多属性文档子空间主题建模方法的使用场景.

5.1.1 游戏评论数据集案例

该案例使用的数据集[62]由 2 986 个样本构成，每个样本对应一个游戏评论. 每个样本包含一个评论文本内容以及本文挑选的 5 个属性，分别是评论的游戏 ID、是否推荐该游戏、用户是否购买该游戏、用户游戏总时长和用户评论时游戏时长. 本文利用这 5 个属性和评论文本作为样本数据，对其进行语义模式挖掘.

首先使用游戏 ID 属性对初始数据进行重构得到 10 个子空间，表征重构得到的 2D 投影如图 17 所示. 其中，相同游戏 ID 的评论能够形成聚簇，分析人员选择聚簇进行主题提取，并选择是否推荐属性以判断游戏的受欢迎程度. 可以看出，在分析人员圈选的主题中, rockstar 主题和 region lock China 主题代表的《侠盗猎车手 5》(grand theft auto V, GTAV)和绝地求生(player unknowns battle grounds, PUBG)这 2 款游戏的不推荐评论数量高于推荐评论数量，用户表达了对游戏公司和作弊现象的不满，游戏中大量不和谐内容同样也引起诸多争议；泰拉瑞亚(Terraria)的推荐游戏评论远高于不推荐的评论数量，在 Steam 上获得了高达 98%的好评率，这个事实与分析人员分析的结果一致.

SP 中存在一个由一些没有分开，颜色杂乱的点集合成的聚簇，如图 17 中 SP 所示，该主题(crust, lattice, sugar, brown, flour)包含一些菜谱评论，是用户推荐游戏后评论与游戏内容无关的恶作剧，也是获得点赞的常用方法.

在游戏 ID 为《收获日 2》(pay day 2, PD2)的子空间节点，根据是否推荐游戏继续添加属性子空间，用来在 SP 中探索用户不推荐 PD2 的原因，发现用户对游戏付费和官方不作为现象表达不满. 通过上述分析，分析人员能够更好地识别值得信赖的游戏开发者/发行商.

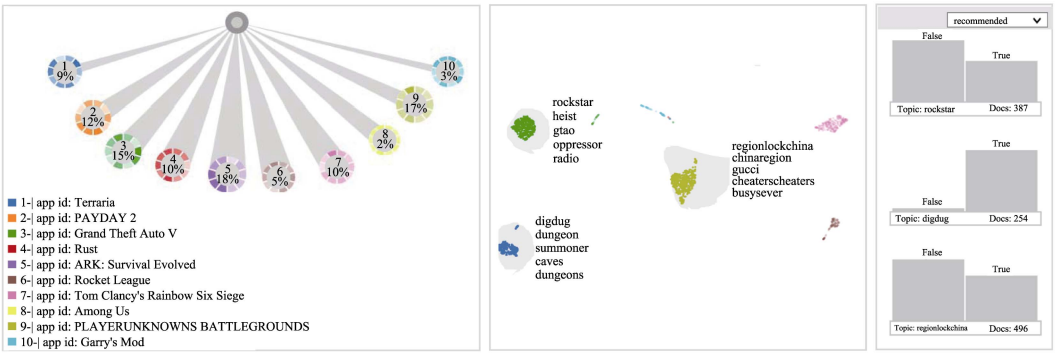


图 17　游戏评论数据集[62]案例

5.1.2　美国新闻数据集案例

该案例使用的数据集是美国出版社发布的新闻数据集[63]，本文选择该数据集中 6 家出版社出版的文章，共包含 1 000 条样本数据. 每个样本由对应一篇文章的文本内容以及 3 个属性(出版社、年份和月份)组成，该数据集中，文章出版时间主要是 2016-01—2017-07，少部分是 2015 年.

分析人员首先使用年份将初始样本进行重构，得到 3 个子空间(年份分别为 2015, 2016, 2017). 年份为 2017 的子空间主题涉及特朗普通俄事件，该主题的月份的讨论度如图 18 中 EM 所示，符合事件发展.

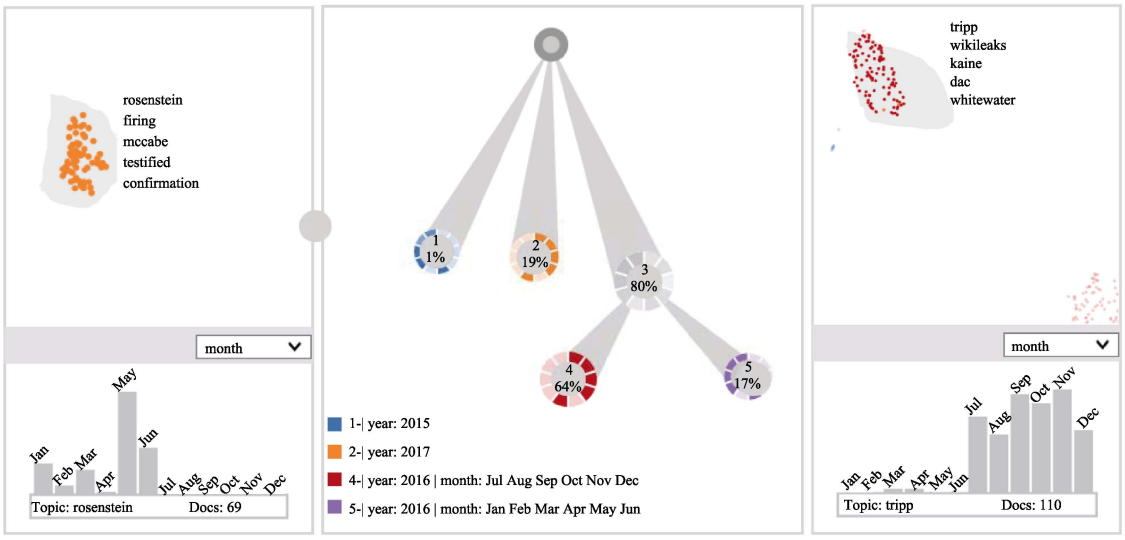


图 18　美国新闻数据集[63]案例

年份为 2016 子空间主题涉及希拉里邮件门事件，分析人员从时间发展的视角探索这些新闻事件. 将该子空间节点按照月份分为 2 组(2016-01—2016-06 和 2016-07—2016-12)进行重构，得到新的子空间，如图 18 中 SP 所示. 希拉里邮件门新闻的时间分布在 2016-07—2016-12，如图 18 中 TE 所示. 主题分布符合事件发展的时间线.

分析人员选择出版社属性对年份为 2016 的子空间节点进行重构，探索不同出版社的政治立场. starr, operate, gormley, whitewater, jury 等主题是华盛顿邮报对邮件门事件未起诉希拉里的情况；source, saudi, muslim, breitbart, world 主题大部分新闻来自布赖特巴特出版社，该出版社的新闻主要抨击希拉里. 华盛顿邮报是左翼出版社，布赖特巴特是极右翼出版社，而希拉里属于左翼政党，本文得到的主题符合出版社的政治立场.

5.1.3　特朗普大选数据集案例

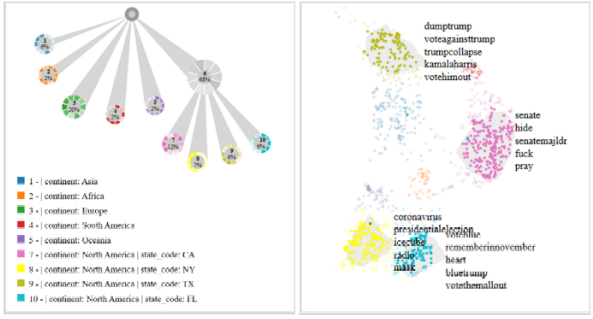
该案例使用了 2020 年美国竞选中有关特朗普

内容的推特数据集[64]. 本文只选取英文推文并随机提取该数据集 3 000 条样本数据，每条数据包含一个文本内容以及时间、大洲和美国州代号 3 个属性，本文选择加利福尼亚(CA)，纽约(NY)，得克萨斯(TX)和佛罗里达(FL)这 4 个州.

分析人员首先使用大洲对初始数据进行重构，如图 19a 所示，得到 6 个子空间，分别代表亚洲、非洲、欧洲、南美洲、大洋洲和北美洲，对 SP 的观察和主题建模，得到以下信息：非洲、亚洲和大洋洲的人们共同关心大选和经济情况，但是对特朗普的参选没有明确的态度；欧洲和南美洲的人们对美国大选的热度不高且没有明确态度，他们更多关注社会民生；在北美洲，人们对特朗普的不满较其他洲更强烈. 为了更好地了解美国不同州之间对特朗普选举的看法，分析人员选择北美洲节点，筛选 4 个州和其他州，以州属性进行重构，得到新的 5 个子空间节点，分布代表北美洲下的佛罗里达、得克萨斯、纽约、加利福尼亚和其他州，如图 19b 所示.



a. 初始子空间



b. 多属性子空间

图 19 美国大选数据集[64]案例

分析人员观察重构子空间，得出以下结论：佛罗里达州是关键的摇摆州；分布在加利福尼亚州的主题关键词表达了对特朗普的反对，说明该州是倾向于民主党的蓝州；纽约州人民更多是讨论选举和社会现象，难以通过主题判断倾向；得克萨斯州普遍被认为是传统红州，近年来该州被列入了摇摆州的行列，与主题中反对特朗普的倾向一致.

5.2 用户实验

通过用户实验验证本文方法的有效性，揭示用户在使用该方法时面临的挑战和优势.

5.2.1 实验设置

(1) 参与者和数据. 本文招募 20 名计算机科学专业和参加自然语言处理课程的研究生作为参与者，包括 8 名男性和 12 名女性，年龄 22~24 岁，中位数为 23 岁；在实验开始前，给每位参与者提供介绍算法原理的教程和自由进行语义模式探索的培训. 评估使用第 5.1.3 节描述的美国大选数据集.

(2) 实验方法. 本文将参与者随机地分成 2 组，每组 10 人，记为 G_A 和 G_B. 其中，G_A 参与者可以不受任何限制地使用本文系统；而当 G_B 参与者使用该系统时，标签重构功能对他们不可见(G_B 可以使用 EM 的功能，如编辑标签和过滤数据).

(3) 任务. 每个参与者使用本文系统挖掘美国大选数据集，探索数据集，找到特定时间内出现的 2 个有趣的语义模式. 实验的目的是期望参与者能够提供对选举过程中动态变化的深刻见解.

(4) 过程. 本文使用屏幕记录软件记录参与者在实验过程中的操作情况. 每个参与者按照如下顺序进行实验：a. 创建子空间；b. 调整 CLR 以生成重构 SP；c. 查看主题和分布；d. 解释语义模式；e. 保存结果. 通过问卷调查收集参与者对本文系统的主观满意度，收集参与者的反馈.

5.2.2 客观表现

本实验收集所有参与者的视频，并手动提取他们的操作序列和每次操作的持续时间后进行评估. 图 20 所示为所有参与者的操作序列和实验结果. 可以看出，G_A 参与者($P_1 \sim P_8$)比 G_B 参与者($P_9 \sim P_{16}$)进行的操作更少，花费的时间也更少. 对实验结果进行 ANOVA 方差分析的结果表明，2 组之间的差异是显著的. 得出以下结论:

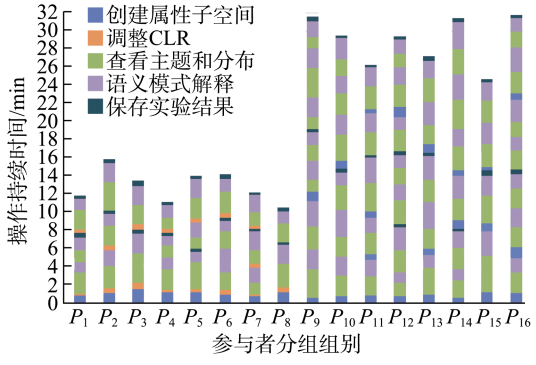


图 20 用户研究实验结果

(1) 调查结果符合本文的预期. 文档重构促使形成子空间的主题模式，从而促进了 G_A 参与者的模式选择;

(2) 唯一的超参数 CLR 不影响本文方法的可用性;

(3) 复杂的属性子空间的语义探索通常需要更多的操作，与较长的完成时间和复杂的操作序列相一致;

(4) 本文系统得到的语义模式更准确且易于理解;

(5) 参与者认为，重构后的主题能更精准地反映子空间语义模式，本文的分析工具有助于对主题的解释;

(6) 用户的主观感知在语义探索中仍然起着重要的作用.

5.2.3　主观评价

本文通过问卷的形式收集 G_A 每个参与者的反馈，参与者使用 7 分制李克特量表[65]对 7 个问题打分，问题如下:

Q1. 在属性树上选择重构子空间操作简单且灵活?

Q2. 重构的主题满足你期待的子空间的语义?

Q3. SP 能够提供分析语义模式的帮助?

Q4. TE 提供的精细语义对于理解主题有帮助?

Q5. 在指定属性的情况下，重构帮助你得到了有趣的语义模式?

Q6. 自由探索阶段，本文系统能够帮助快速地理解数据集?

Q7. 你在探索语义过程中对产生的语义模式感到困惑了吗?

每个问题总分为 7 分，分数越高，表示参与者对这个方面就越满意. 调查问卷统计结果如图 21 所示. 可以看出，大多数参与者对本文提出的语义探索工具给出了较高的满意度评分(平均 36.50 分)，反映了该工具的有效性和用户友好性; Q3 和 Q4 的得分相对较高，表明参与者普遍认可 SP 和 TE 在查看和理解文档语义模式方面的有效性; 本文工具在发现有趣模式(Q5)和帮助理解数据集(Q6)方面表现突出，这不仅证明了本文系统在已知语义模式发现中的有效性，还展示了其在发现新颖、有趣语义模式方面的潜在价值，更重要的是，该系统降低了语义模式探索的用户门槛，使得即使是未了解数据集的用户，在经过简单的培训后，也能快速、全面地理解文档语义; 用户并未明显地感知到

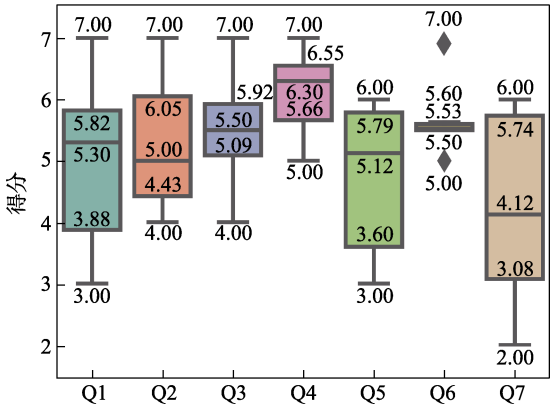


图 21　调查问卷统计结果

同时重构多个属性子空间可能带来的困惑，进一步证明了本文工具的易用性和稳定性.

综上所述，调查问卷的结果表明，本文提出的语义探索工具在帮助用户理解和探索语义模式方面表现良好，具有较高的执行效率和有效性. 未来将基于这些反馈进行进一步的优化和拓展.

5.3　量化实验

通过 2 组实验，验证本文方法的性能: 第 1 组实验评估本文方法在获取主题-词分布时的主题质量; 第 2 组实验用于评估模型的性能.

5.3.1　主题模型评价指标

本文将主题质量的总体度量定义为其主题多样性(topic diversity, TD)和主题一致性. Mimno 等[66]提出使用主题一致性评价主题可解释性. 本文实验采用一致性值(coherence value, CV)指标计算一致性评分，计算公式为

$$\mathrm{CV} = \sum v \Big/ \mathrm{C}_n^2 \tag{5}$$

其中，v 用于衡量主题中的 2 个不同单词的共现频率，通常使用点互信息(pointwise mutual information, PMI)或其他度量表示 2 个单词之间的相关性.

TD 可以理解为在 T 组主题中，排名前 N 个单词中不同单词的比例，该指标的目的是衡量主题之间的差异性，帮助评估主题模型的质量. TD 的计算公式为

$$\mathrm{TD} = \frac{K}{T \times N} \tag{6}$$

其中，K 表示在排名前 N 个关键词中出现的独立关键词的数量.

在游戏评论数据集、美国新闻数据集和特朗普大选数据集上，对比 LDA[6], NMF[8], BERTopic[16]和本文方法的评价指标，结果如表 1 所示. 其中，本文方法的主题是分析人员随机选择重构属性后在 SP 中套索选择得到的. 可以看出，本文方法在

TD 方面表现十分出色，明显优于 3 种对比方法的得分，说明其生成的主题中更好地包含了数据集中的多样化信息；但是，本文方法在主题一致性上的表现略低于 BERTopic 和 NMF 方法，这是因为表征重构过程在尝试覆盖数据集中更多信息、生成更多样化主题的同时，也在一定程度上牺牲了主题内部的凝聚力，由于该方法引入子空间属性的影响，使得原本在语义上相近但属于不同子空间的文档可能在重构后的潜空间中被分散，影响了主题内部的一致性；游戏评论数据集中包含大量玩家之间的术语和特定领域的语言，美国新闻数据集涉及广泛的主题，包括政治、经济等，本文方法采用特定的 SP 方法和套索选择策略，能更适应不同领域和语义复杂性的数据，有效地捕捉了数据集中的多样化信息，进而生成更加精细和多样化的主题，在主题建模上具有较好的泛用性和灵活性.

表 1　在 3 个数据集上 4 种方法主题模型评价指标对比

数据集	主题一致性				TD			
	LDA[6]	BERTopic[16]	NMF[8]	本文	LDA[6]	BERTopic[16]	NMF[8]	本文
游戏评论	0.420	**0.690**	0.462	0.436	0.580	0.740	0.540	**0.890**
美国新闻	0.392	**0.598**	0.527	0.566	0.560	0.930	0.740	**0.950**
特朗普大选	0.355	0.366	**0.483**	0.414	0.440	**0.940**	0.710	0.930

注. 粗体表示最优值.

5.3.2　模型性能

为了评估本文系统中子空间主题模型的性能，分别在BBC新闻数据集[67]、航空公司评论数据集[68]、希拉里-特朗普推特发文数据集[69]、新闻出版社数据集[63]和仇恨言论推文数据集[70]上，分别使用 GPU 和 CPU 对一轮子空间主题建模的执行时间和内存开销进行实验，结果如表 2 所示. 其中，每个数据集具有不同的文档数量和属性，每个文档样本具有语义表征和属性标签；实验中使用的服务器设置为 12th Gen Intel(R) Core(TM) i5-12500, 16 GB, RTX2060. 可以看出，本文方法的执行时间和内存开销(x 轴方向)随着数据集文档数量(y 轴方向)的增加而逐渐增加，但是，即使是在 10 000 个文档数量下，其执行时间也仅需 4.0 s，能够满足可视化系统的实时交互要求.

表 2　在 5 个数据集上芝加哥犯罪数据模型效率评估结果对比

数据集	文档数量	执行时间/s		内存开销/MB	
		GPU	CPU	GPU	CPU
BBC 新闻[67]	1 490	1.400	3.000	1 020.300	42.900
航空公司评论[68]	3 701	1.600	18.300	1 029.300	46.100
希拉里-特朗普推特发文[69]	6 444	2.700	53.700	1 039.700	49.400
新闻出版社[63]	10 000	4.000	130.500	1 060.500	184.800
仇恨言论推文[70]	20 000	12.100	562.100	1 162.800	328.300

6　结　语

本文针对语义模式的探索设计了一个表征重构网络，将潜空间中的语义坐标映射为包含语义和属性信息的潜向量，通过主题模型识别子空间主题；采用 HITL 的语义模式可视分析方法，辅以本文构建的可视化系统，支持用户调整可视化和交互组件，以及量化指标、观测数据分布、选择表征重构的目标属性和感兴趣的子空间，以获得特定语义模式.

在3个真实数据集上对本文方法进行评估，验证了该方法和本文系统的有效性；在 20 位被试进行的用户研究中，参与者的使用情况和反馈显示，本文方法提高了分析人员挖掘语义模式的效率，有助于更深入地理解数据集的语义；最后通过2组量化实验，验证了该方法的主题质量和模型性能.

未来将从 2 个维度扩展本文工作. (1) 本文方法虽然使得用户能够通过观察数据分布结合主题提取获得语义模式，但会产生降维导致的数据信息的丢失和维度诅咒等问题. 今后可以利用可视化隐喻[71-72]设置不同图形代替聚簇数量和大小，进一步探索高维投影结构和结构化的可视化形式，提高用户语义模式挖掘的分析速度. (2) 本文采用 TE-IDF 的方法计算主题词，今后考虑将不同的主题提取方法与本文系统结合，让用户能够根据具体需求选用合适的方法.

参考文献(References):

[1] Li J, Chen S M, Andrienko G L, *et al*. Visual exploration of spatial and temporal variations of tweet topic popularity[C]//Proceedings of the EuroVis Workshop on Visual Analytics. Aire-la-Ville: Eurographics Association Press, 2018: 7-11

[2] Felix C, Pandey A V, Bertini E, *et al*. RevEx: visual investigative

journalism with a million healthcare reviews[OL]. [2024-03-01]. http://cj2015.brown.columbia.edu/papers/revex.pdf

[3] Li J, Chen S M, Chen W, *et al*. Semantics-space-time cube: a conceptual framework for systematic analysis of texts in space and time[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(4): 1789-1806

[4] Aggarwal C C, Zhai C X. A survey of text classification algorithms[M] //Aggarwal C C, Zhai C X. Mining Text Data. New York: Springer, 2012: 163-222

[5] Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 1972, 28(5): 11-21

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

[7] Deerwester S, Dumais S T, Furnas G W, *et al*. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407

[8] Xu W, Liu X, Gong Y H. Document clustering based on non-negative matrix factorization[C] //Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2003: 267-273

[9] Blei D M, Jordan M I, Griffiths T L, *et al*. Hierarchical topic models and the nested Chinese restaurant process[C] //Proceedings of the 16th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2003: 17-24

[10] The Y W, Jordan M I, Beal M J, *et al*. Sharing clusters among related groups: hierarchical Dirichlet processes[C] //Proceedings of the 17th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2004: 1385-1392

[11] Lacoste-Julien S, Sha F, Jordan M I. DiscLDA: discriminative learning for dimensionality reduction and classification[C] //Proceedings of the 21st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2008: 897-904

[12] Blei D M, Mcauliffe J D. Supervised topic models[C] //Proceedings of the 20th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2007: 121-128

[13] Liu Y, Liu Z Y, Chua T S, *et al*. Topical word embeddings[C] //Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015: 2418-2424

[14] Qiang J P, Chen P, Wang T, *et al*. Topic modeling over short texts by incorporating word embeddings[C] //Proceedings of the 21st Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Heidelberg: Springer, 2017: 363-374

[15] Angelov D. Top2Vec: distributed representations of topics[OL]. [2024-03-01]. https://arxiv.org/abs/2008.09470

[16] Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure[OL]. [2024-03-01]. https://arxiv.org/abs/2203.05794

[17] Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts[J]. Frontiers in Sociology, 2022, 7: Article No.886498

[18] Song X L, Wang X T, Hu X H. Semantic pattern mining for text mining[C] //Proceedings of the IEEE International Conference on Big Data (Big Data). Los Alamitos: IEEE Computer Society Press, 2016: 150-155

[19] Naïm H, Aznag M, Durand N, *et al*. Semantic pattern mining based web service recommendation[C] //Proceedings of the 14th International Conference on Service-Oriented Computing. Heidelberg: Springer, 2016: 417-432

[20] Zhang C, Han J W, Shou L D, *et al*. Splitter: mining fine-grained sequential patterns in semantic trajectories[J]. Proceedings of the VLDB Endowment, 2014, 7(9): 769-780

[21] Cao N, Lin Y R, Sun X H, *et al*. Whisper: tracing the spatio-temporal process of information diffusion in real time[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2649-2658

[22] Hu M D, Wongsuphasawat K, Stasko J. Visualizing social media content with sententree[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 621-630

[23] Alexander E, Kohlmann J, Valenza R, *et al*. Serendip: topic model-driven visual exploration of text corpora[C] //Proceedings of the IEEE Conference on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2014: 173-182

[24] Dou W W, Wang X Y, Chang R, *et al*. ParallelTopics: a probabilistic approach to exploring document collections[C] //Proceedings of the IEEE Conference on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2011: 231-240

[25] Cui W W, Liu S X, Wu Z F, *et al*. How hierarchical topics evolve in large text corpora[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 2281-2290

[26] Liu S X, Wang X T, Chen J F, *et al*. TopicPanorama: a full picture of relevant topics[C] //Proceedings of the IEEE Conference on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2014: 183-192

[27] Kim M, Kang K, Park D, *et al*. TopicLens: efficient multi-level visual topic exploration of large-scale document collections[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 151-160

[28] Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model[C] //Proceedings of the 13th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2000: 893-899

[29] Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality[C] //Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2013: 3111-3119

[30] Le Q, Mikolov T. Distributed representations of sentences and documents[OL]. [2024-03-01]. https://doi.org/10.48550/arXiv.1405.4053

[31] Lau J H, Baldwin T. An empirical evaluation of doc2vec with practical insights into document embedding generation[OL]. [2024-03-01]. https://doi.org/10.48550/arXiv.1607.05368

[32] Devlin J, Chang M W, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding[C] //Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Minneapolis: Association for Computational Linguistics, 2019: 4171-4186

[33] 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

[34] Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019: 3982-3992

[35] Chicco D. Siamese neural networks: an overview[M] //Cartwright H. Artificial Neural Networks. New York: Springer, 2021: 73-94

[36] McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction[OL]. [2024-03-01]. https://arxiv.org/abs/1802.03426
[37] Hu Y N, Boyd-Graber J, Satinoff B, *et al*. Interactive topic modeling[J]. Machine Learning, 2014, 95(3): 423-469
[38] 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
[39] Park D, Kim S, Lee J, *et al*. ConceptVector: text visual analytics via interactive lexicon building using word embedding[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 361-370
[40] 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
[41] 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, 2019, 26(1): 1001-1011
[42] 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
[43] Smith A, Kumar V, Boyd-Graber J, *et al*. Closing the loop: user-centered design and evaluation of a human-in-the-loop topic modeling system[C] //Proceedings of the 23rd International Conference on Intelligent User Interfaces. New York: ACM Press, 2018: 293-304
[44] Beeferman D, Jiang H. Topic-time heatmaps for human-in-the-loop topic detection and tracking[OL]. [2024-03-01]. https://arxiv.org/pdf/2110.07337
[45] Khan M H U R R, Wakabayashi K. Reducing human effort in keyphrase-based human-in-the-loop topic models: a method for keyphrase recommendations[C] //Proceedings of the 25th International Conference on Information Integration and Web Intelligence. Heidelberg: Springer, 2023: 233-248
[46] Johnson B, Shneiderman B. Tree-maps: a space-filling approach to the visualization of hierarchical information structures[C] //Proceedings of the Visualization. Los Alamitos: IEEE Computer Society Press, 1991: 284-291
[47] Long L K, Hui L C, Fook G Y, *et al*. A study on the effectiveness of tree-maps as tree visualization techniques[J]. Procedia Computer Science, 2017, 124: 108-115
[48] Barlow T, Neville P. A comparison of 2-D visualizations of hierarchies[C] //Proceedings of the IEEE Symposium on Information Visualization. Los Alamitos: IEEE Computer Society Press, 2001: 131-138
[49] Cai G, Sun F, Sha Y Z. Interactive visualization for topic model curation[OL]. [2024-03-01]. https://ceur-ws.org/Vol-2068/esida5.pdf
[50] Jiang X Y, Zhang J W. A text visualization method for cross-domain research topic mining[J]. Journal of Visualization, 2016, 19(3): 561-576
[51] Shao L, Schleicher T, Behrisch M, *et al*. Guiding the exploration of scatter plot data using motif-based interest measures[J]. Journal of Visual Languages & Computing, 2016, 36: 1-12
[52] Chen H S, Wang X M, Pan S R, *et al*. Identify topic relations in scientific literature using topic modeling[J]. IEEE Transactions on Engineering Management, 2021, 68(5): 1232-1244
[53] 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
[54] Xie T, Qin P, Zhu L B. Study on the topic mining and dynamic visualization in view of LDA model[J]. Modern Applied Science, 2018, 13(1): Article No.204
[55] Zou C Y, Hou D Q. LDA analyzer: a tool for exploring topic models[C] //Proceedings of the IEEE International Conference on Software Maintenance and Evolution. Los Alamitos: IEEE Computer Society Press, 2014: 593-596
[56] Roberts J C. State of the art: coordinated & multiple views in exploratory visualization[C] //Proceedings of the 5th International Conference on Coordinated and Multiple Views in Exploratory Visualization. Los Alamitos: IEEE Computer Society Press, 2007: 61-71
[57] Xie P, Tao W Y, Li J, *et al*. Exploring multi-dimensional data via subset embedding[J]. Computer Graphics Forum, 2021, 40(3): 75-86
[58] Li J, Zhou C Q. Incorporation of human knowledge into data embeddings to improve pattern significance and interpretability[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 29(1): 723-733
[59] MacQueen J. Some methods of classification and analysis of multivariate observations[C] //Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967: 281-297
[60] Bostock M, Ogievetsky V, Heer J. D^3 data-driven documents[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2301-2309
[61] Abadi M, Agarwal A, Barham P, *et al*. TensorFlow: large-scale machine learning on heterogeneous distributed systems[OL]. [2024-03-01]. https://arxiv.org/abs/1603.04467
[62] Marko M. Steam reviews dataset 2021[EB/OL]. [2024-03-01]. https://www.kaggle.com/datasets/najzeko/steam-reviews-2021
[63] All the news[EB/OL]. [2024-03-01]. https://components.one/datasets/all-the-news-articles-dataset/
[64] Manch H. US election 2020 tweets[EB/OL]. [2024-03-01]. https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets
[65] Clark L A, Watson D. Constructing validity: basic issues in objective scale development[M] //Kazdin A E. Methodological Issues and Strategies in Clinical Research. 4th ed. Washington: American Psychological Association, 2016: 187-203
[66] Mimno D, Wallach H, Talley E, *et al*. Optimizing semantic coherence in topic models[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh: ACL, 2011: 262-272
[67] BBC news classification[EB/OL]. [2024-03-01]. https://www.kaggle.com/competitions/learn-ai-bbc/data
[68] Anshul C, Raisinghani M. Airline reviews[EB/OL]. [2024-03-01]. https://www.kaggle.com/datasets/chaudharyanshul/airline-reviews
[69] Hamner B. Hillary Clinton and Donald trump tweets[EB/OL]. [2024-03-01]. https://www.kaggle.com/datasets/benhamner/clinton-trump-tweets
[70] The Devastator. Hate speech and offensive language detection[EB/OL]. [2024-03-01]. https://www.kaggle.com/datasets/thedevastator/hate-speech-and-offensive-language-detection
[71] Chen S, Li S H, Chen S M, *et al*. R-Map: a map metaphor for visualizing information reposting process in social media[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 1204-1214
[72] Yang H, Li J, Chen S M. TopicRefiner: coherence-guided steerable LDA for visual topic enhancement[J]. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(8): 4542- 4557