词云可视化综述

包琛, 汪云海*

(山东大学计算机科学与技术学院 青岛 266237)
(wang.yh@sdu.edu.cn)

摘 要: 词云是一种近年来颇为流行的文本可视化方式, 它提取出文本中的关键词并在二维空间上美观地排布, 通常用于展示文本内容、辅助文本分析以及吸引读者阅读等. 从视觉编码、布局方法和交互方式这 3 个方面介绍词云的设计空间; 将现有的词云设计分为语义词云、形状词云、可编辑词云和多文档词云 4 类进行概括, 并总结了目前对于词云进行实验评价的若干工作; 最后分别从语义词云、形状词云、多文档词云和中文词云 4 个方面分析了词云可视化领域面临的挑战, 并对未来工作进行了展望.

关键词: 词云; 标签云; 文本可视化; 语义词云; 形状词云; 多文档文本; 文本分析
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2021.18811

A Survey of Word Cloud Visualization

Bao Chen and Wang Yunhai*

(*School of Computer Science and Technology, Shandong University, Qingdao* 266237)

Abstract: Word cloud is a popular text visualization technique that extracts keywords from text and displays them on the 2D space aesthetically. Word cloud is often used to display contents, aid text analysis and attract readers. In this work, the design space of word cloud is introduced from three aspects: visual encoding, layout and interaction. Then current word cloud design researches are summarized by four categories: semantic word clouds, shape-constrained word clouds, interactive word clouds and multi-document word clouds. Several works related to word cloud evaluation are also concluded. Finally, research challenges in semantic word clouds, shape-constrained word clouds, multi-document word clouds and Chinese word clouds, and suggest future work of word cloud visualization are discussed.

Key words: word cloud; tag cloud; text visualization; semantic word clouds; shape-constrained word clouds; multi-document text; text analysis

词云是一种对文本进行总结概括的可视化方法, 通过提取关键词在二维空间中排布, 词云以友好的方式向人们展示了文本的主要内容. 通常, 词云中的单词大小由单词在文本中的出现频率映射而来, 直观地表达单词的重要程度. 词云因其美学上的优越性以及其简洁易懂的特性广受人们喜爱, 尤其是在广告、新闻、教育和出版等行业中得到了非常广泛的应用. 例如, 在照片分享社区网站, 图片标签以词云的方式呈现, 起到了展示图片热度和网页导航的作用[1]; 新闻记者热衷于使用词云作为吸引读者注意力的头条图片, 内容包含新闻事件的对比和人物事迹的概括等[2]; 普通用户使用词云制作个人简历和邮件个性签名[3], 印制纪念品和海报; 在教育行业, 词云还可被用于概括课程

收稿日期: 2020-11-24; 修回日期: 2021-02-18. **基金项目**: 国家自然科学基金(61772315, 61861136012). **包琛**(1995—), 女, 硕士研究生, 主要研究方向为信息可视化; **汪云海**(1984—), 男, 博士, 教授, 博士生导师, 论文通讯作者, 主要研究方向为数据可视化、人机交互.

主要内容或辅助科学研究教学[4].

随着近年词云被大量使用, 相关的研究和设计也逐渐丰富. 研究者在传统视觉编码和布局方法的基础上, 不断地试图增强词云的表达能力, 逐渐发展出多样的词云外观形式, 大大丰富了词云的内涵. 此外, 尽管已经有很多研究证明词云不能成为一个有效的数据分析工具, 鉴于词云的受欢迎程度, 许多可视化研究者开始从实验的角度试图评价词云设计的好坏, 探索词云的设计空间. 本文通过分类陈述的方式, 总结了词云自诞生以来在相关研究领域的发展和演变, 意在为词云研究者提供一个概括的了解, 同时启发设计者选择适合自己的词云展示方案.

本文首先对词云的基本特性和设计空间进行了介绍; 其次阐述4类词云可视化设计的特征和代表性成果; 然后概述当前词云实验评价的相关研究; 最后总结并探讨了词云未来可能的发展方向及面临的挑战. 鉴于目前网络上有层出不穷的词云自动生成工具, 本文将讨论的范围限制在目前已发表成文并且明确说明技术方法的词云研究工作之中. 文中会涉及少量商业化词云工具, 本文将仅描述其效果, 避免探讨背后的技术原理.

1　词云简介

1.1　词云的设计空间

词云又称为标签云, 早期用于展示文档或数据的标签. Rivadeneira 等[5]提出, 构建词云主要依赖 2 种类型的特征: (1) 文本特征, 包括字体的粗细、大小、颜色; (2) 单词排布的特征, 包括排序、聚类、空间布局. 之后, Felix 等[6]将视觉编码和布局方法作为词云的 2 个最关键的可视参数, 其中, 视觉编码包括字体通道(颜色、大小等)以及附加符号通道(添加柱、圆等).

Bateman 等[7]将词云的可视变量划分得更详细, 考虑了英文字母中不同字母占用的像素数量不同、字母宽度不同等, 但实验证明这些因素并非影响人们对于词云感知的主要因素.

随着词云的普及, 用户对词云定制化的需求逐渐增长, 出现了一些对已生成词云进行编辑交互的工作[8-10]. 参考前人的总结方式并结合当前词云的研究进展, 本文将词云的设计空间分为如图 1 所示视觉编码、布局方法与交互方式 3 个部分, 并依据目前已有的词云工作列出了常见的几种参数.

1.1.1　视觉编码

词云使用的主要视觉编码通道是文字本身, 其中用字体大小表示单词重要性(通常为词频)是最常见的编码方式. 除此之外, 也有一些工作使用颜色、透明度等作为词频的冗余编码(指对同一维度同时使用多个通道进行编码), 或者表示除词频外的其他信息. 例如, 在多文档词云中, 可以使用颜色区分从上一个时间步到当前时间步单词发生的变化[11], 或者使用透明度表示单词的逆文档频率(inverse document frequency, IDF)值(即包含当前单词的文档数比例)[12].

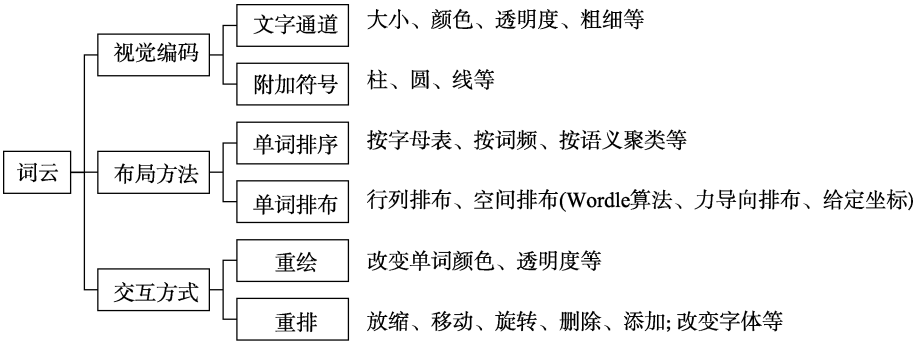

图 1　词云的设计空间

另外, 尽管研究人员尝试过为词云添加其他可视化图形, 但在不破坏词云美观性的前提下使用附加符号是比较困难的事情, 目前常见的往往是通过添加折线表示词频变化趋势. Collins 等[13]提出将平行坐标系和标签云结合的方式, 数据在时序上发生的变化是通过平行坐标系中的连线来展示的. SparkClouds[14]给标签云中的每个单词添加迷你趋势线用以展示时序数据. 文献[15]将直方图与标签云结合起来, 以展示随时间变化的单词共现关系, 用户点击某个单词进行交互时, 系统会高亮显示与之共现的单词以及共现的时间段.

文献[16]在尝试给标签云添加时序信息表达

时，考虑了上述的 2 个方面：对于文字通道本身，使用文字的亮度、大小、变形程度和透明度表现频率随时间的变化；对于附加图形方面，使用颜色背景展示标签发生的变化，带颜色的线段展示标签频率随时间的变化，带颜色的日历表和圆形图符展示重复出现或循环的数据规律．该工作比较综合地展示了不同通道的效果，给用户提供了多样的选择．

然而，为词云添加过于复杂的视觉编码方式可能会损害词云自身易读性[16]，给人们带来认知上的额外负担．因此，设计词云时如果要采用非常见的编码通道或添加较复杂的可视化图形，应该慎之又慎．

1.1.2 布局方法

词云至今已经发展出了多种多样的布局方法，人们可以使用不同的排序与排布方法展示单词，可以说，布局是词云的核心．早期比较常用的是水平竖直规则排布的行列式标签词云，这个阶段较多使用字母表顺序对单词排序；而经典的 Wordle 算法[3]诞生并流行至今，其排序方法往往和词频或其他计算单词重要性的表示方法有关；此外，力导向排布在语义词云中也有比较广泛的使用．鉴于这一部分的重要性，下面详细介绍这 3 种最常见的单词排布方法．

(1) 行列排布．又称为水平竖直排布，就是将单词在画布上从左到右或从上到下对齐排列，是一种常见的词云排布样式．单词通常按照字母表顺序或按照它们的权重排序．Parallel tag clouds[13]和 SparkClouds[14]是典型的行列排布单词的词云．这种布局方法的优点是结构清晰，一目了然，有研究表明这种布局方式往往有利于人们完成大小判断和关键词检索等底层任务[6]．另外，也有语义词云的研究人员称这种布局方式能够帮助人们完成提取概括文章主题这种高层感知任务[17]．这种排布方式被诟病之处是相对比较死板，美观性较差．与之相对的是空间排布方式，主要有 Wordle 算法和力导向排布 2 种．空间排布的单词不再追求对齐工整，视觉上更具吸引力．

(2) Wordle 算法．是词云的一种经典排布，它生成的词云自然、美观且紧凑，如图 2 所示．Feiberg[3]详细介绍了 Wordle 的起源、单词排布策略、碰撞检测及其效率优化方法和代码实现等．算法核心就是将单词按照权重由大到小排序，然后从画布中间开始按照顺序逐个摆放，要摆放的单词需要与已放置的单词之间进行碰撞检测，如果发现与已放置单词产生交叠，要摆放的单词沿着阿基米德螺旋线的路径往外移动一步；重复进行碰撞检测和移动，直到没有交叠将单词放下为止；重复以上过程直至摆放好所有单词．这种贪婪算法尽管复杂度较高，但是因为其生成结果的高度美观性，目前非常流行．Feiberg[3]还讨论了 Wordle 算法面向不同用户和不同任务的优劣．因为 Wordle 网站①自诞生以来广受欢迎，拥有成千上万的用户，Viegas 等[2]收集了几年内人们使用 Wordle 算法进行的创作结果并发表了针对用户使用的调研[2]．除了圆形螺旋线之外，矩形螺旋线也是一种常见的变体．此外，Wang 等[18]提出一种改造传统螺旋线使其适应任意形状的方法，进一步丰富了螺旋线布局的内涵．

图 2 Wordle 算法生成的布局[2]

(3) 力导向排布．如果将单词看做图中的点，并为点与点之间添加边，就可以使用力导向模型对词云中的单词进行布局．例如，基于刚体动力学系统[19]的方法将每个单词看做一个有体积的刚体，充分利用了力之间的吸引和排斥作用，将单词之间的距离控制在合适大小的同时，避免了单词之间的重叠．这有利于保持词云紧凑且没有重叠的优良特性．同时，由于单词之间的距离可以自然联想到用语义上的距离来替代，因此力导向排布在语义词云的分类下有着比较广泛的应用，并且往往会和降维方法结合，将单词在高维空间的语义关系呈现在二维空间．然而，这种布局方法也有其固有缺陷：该方法有时无法达到预期目标效果[20]；使用力导向排布[21-22]不能保证结果一定会达到预期，因为难以预测复杂的力相互作用的结果，降低了词云生成方法的鲁棒性．

① http://www.wordle.net/

1.1.3　交互方式

Wordle 等在线生成词云的网站可以选择形状词云的外形轮廓, 在生成词云之前设定单词的朝向和词语颜色等. 这些创作工具均提供交互功能为自动一次性生成词云选定参数.

除了这种生成参数设定的情况, 用户可以对已经自动生成的词云中的单词进行再次修改. 用户可通过单击选择词云中的单个单词或框选多个单词进行编辑. 这样的编辑交互主要包括 2 类: 重绘类操作指的是改变单词的颜色和透明度等外观, 不会破坏整体布局; 重排类操作包括对单词的放缩、移动、删除和添加等编辑操作, 或者改变单词的字体等. 重排类操作可能会破坏布局原有的紧凑度和无重叠特性, 进而需要进行重新布局. 每次编辑操作后对未编辑的单词重新运行 Wordle 算法, 可以重新得到美观的布局, 但是这种方式破坏了操作前后单词的位置, 给用户对词云的控制造成了阻碍. 在保留原始的紧凑度和无重叠的情况下, 使用力导向布局可以实现保持前后一致性的编辑交互. 本文将在第 2.3 节重点介绍会引发重排的可编辑词云.

1.2　词云分类

为了方便给设计师提供快速选择的参考, 本文从功能性的角度将词云分成了 4 大类: 语义词云将自然语言意义上的联系程度转化为二维空间上的展示距离, 意在增强词云的语义能力; 形状词云通过给词云轮廓增加形状限制, 大大提高了美观性, 并且从另一个角度增强了词云的表意能力; 可交互编辑的词云, 随着计算机和触屏设备的广泛使用也成为了一个重要的研究方向; 近年出现了针对多文档数据的词云, 针对时序的或相似文档类型的文本数据设计相应的词云方案, 大规模文档也可使用词云作为主要视图辅助文档的可视分析. 本文将各种类型的词云及其代表性工作[8-14,18-20,22-37]从视觉编码、布局方法和交互方式的角度进行如表 1 所示总结, 并且分类列出了本文中涉及的所有相关文献.

表 1　按功能对词云进行分类

词云类别	代表文献	视觉编码		布局方法		交互方式	其他相关工作
		字体通道	附加符号	单词排序	单词排布	重绘与重排	
语义词云	文献[11]	大小、颜色	线	按语义聚类	力导向	无	文献[24-26]
	文献[22]	大小	无	按语义聚类	力导向	无	
	文献[23]	大小	无	按相关度、按语义聚类	力导向	无	
形状词云	文献[27]	大小	无	按地理信息	地理坐标	无	文献[28-29]
	文献[20]	大小	无	按词频	力导向	无	
	文献[18]	大小	无	按词频	Wordle	重绘、重排	
可编辑词云	文献[8]	大小	无	按词频	Wordle	重绘、重排	
	文献[9]	大小	无	按词频	Wordle	重绘、重排	
	文献[10]	大小	无	按词频	Wordle	重绘、重排	
多文档词云	文献[13]	大小	线	按字母表	行列	无	文献[10-11,19,30-36]
	文献[14]	大小	线	按字母表	行列	无	
	文献[33]	大小	流图	按语义和按词频	Wordle	无	
	文献[12]	大小、颜色、透明度	无	随机	Wordle	无	
	文献[37]	大小、颜色	无	按语义聚类	力导向	重排	

2　词云的设计方法

2.1　语义词云

语义词云的设计者更看重词云表意的功能, 于是将语义上更相关的单词排布得更加接近, 以期更好地表达词云包含的文本含义.

如图 3 所示, 保持上下文的动态词云(context preserving dynamic word cloud, CPD)[11]从时序文本数据中提取出单词聚类为簇, 并最终形成多帧的关联词云. 其提出了 3 种语义相关性的度量准则, 对应 3 种不同的单词的特征向量表示方式, 有针对性地生成不同风格的语义词云. CPD 的基本方法是对于任意一对单词两两计算向量的余弦值, 余弦值越大, 证明 2 个单词的相似性越高, 由此创建相似度矩阵, 并使用多维标度法[38]将矩阵中包含的单词投影到二维平面, 以获取初始的按照语义聚类的单词位置布局. 后续在各个时间步单独的词云中, 通过在每个时间步构造图, 同时依据

“无重叠、扁平化和紧凑”3 个原则添加带限制的力导向模型，对构建的图施加力的作用，以获得较为紧凑的布局结果. CPD 的作者认为，将单词按照语义组织成簇增强了词云的可读性，有利于用户理解并且跟踪文本内容.

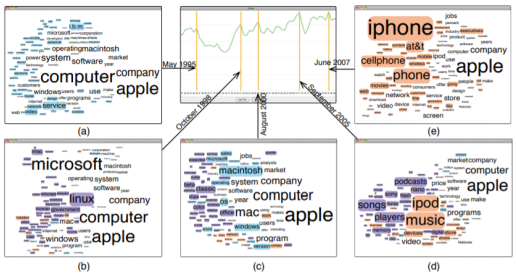

图 3 保持上下文的动态词云[11]

Wu 等[22]认为, CPD 构建的词云存在 2 个问题: (1) 最终的结果不稳定，输入单词的少量变化可能会导致生成完全不同的最终布局; (2) 最终生成的结果可能有非常不规则的外形. 针对此，Wu 等[22]提出了基于图片处理缝隙剪裁方法的词云(seam carving word cloud, SC)，替换掉了 CPD 中的力导向模型，使整体布局紧凑. 原始缝隙裁剪算法首先使用一个能量函数确定布局中的低能量区域，即空白缝隙，然后从左到右或从上到下地裁剪空白. Wu 等[22]针对词云的应用场景做了改进，使用单词的外包围盒作为划分区域的单位，在提高原算法效率的同时，产生了更好的结果.

此外，为了帮助人们比较相关文档，SC 将结果生成的语义词云与气泡集合可视化相结合，以增强语义联系的表达，并最终用并列表格的展示方式辅助用户进行文档之间的比较. Wu 等[22]将 SC 与 CPD 进行比较，说明 SC 在保持外形方面有更好表现，同时也更好地保持了语义联系. 同时，Wu 等[22]也承认无论是 SC 还是 CPD 都更适合文档比较和分析，在美观性和紧凑度方面劣于传统的 Wordle 算法.

ProjCloud[23]是一种用于展示多文档文本的语义词云，如图 4 所示. 其为文档集合中的每篇文档构建一个词云，并且使用降维投影的方法可视化相关文档之间的相似度关系. 由此，文档位置体现了文档之间的相似程度; 将每个词云填充范围限定在一个自动划分或手动划分的多边形内，并保持单词之间的语义关系. ProjCloud 将形状限制和语义保持相结合，以提高了词云的表现力; 同时将空间划分成了多边形区块，具有较高的美观度. Paulovich 等[23]特别提到，尽管其使用了最小二乘投影将多文档向量投影到二维可视空间，事实上，在生成语义词云时，其他降维方法也可能适用. 例如，有的语义词云[24]使用 *t*-SNE 作为投影的方法，也取得了比较不错的效果.

图 4 ProjCloud[23]示例

ReCloud[25]用来展示用户评价(如餐厅点评)，构建辅助用户决策的语义词云. 其使用自然语言处理的技术生成语法依赖图，以保存用户评价中关键词之间的语义，之后同样使用了力导向的方法优化布局. Xu 等[26]使用词向量表示单词，并根据单词之间的语义相似性构建相关图.

总之，当前语义词云的生成算法设计基本上都是通过将单词语义上的联系转化为二维平面的真实距离表现语义，创新点主要集中在 2 个方面: (1) 衡量单词之间语义关联度的方式; (2) 将度量的关联度转化为二维空间坐标系内的位置的方法. 将高维的单词向量投影之后，往往使用力导向的方法保证整体布局的紧凑性.

2.2 形状词云

带有形状限制的词云不仅非常美观，表意能力也比较强，这是因为形状本身就带有对于文本内容的强烈暗示. 本节内容根据输入数据的不同分为 2 个部分: 首先介绍反映地理数据在地图上的分布的形状词云，这类词云除了形状限制之外，还包含根据地理信息摆放单词的要求，如增加城市名称和城市地理坐标相对应的约束; 之后将探讨更一般情况下不包含地理信息的其他形状的词云.

2.2.1 地理形状词云

地理形状词云中的单词通常为地理名称，其位置需要与坐标相对应. 地理词云(geographical word clouds, GWC)[27]是一种根据地理信息放置单词，并且最终形成模拟地图上的地区形状的词云，如图 5 所示. 算法的输入是分布在地理区域内点的二维坐标，每个点都与一个或多个单词相关联.

GWC 设计了若干准则，包括尽量用一个大单词代替多个小单词、保证足够的单词覆盖率、单词大小与其重要性成比例、单词位置不超越区域边界并同时紧凑减少空白等. 在生成 GWC 时，在不同的准则之间要进行取舍. 例如，为了提高紧凑程度可能会改变单词的比例关系，这种做法有损数据真实性，但最终结果的美观程度较高.

图 5　地理词云[27]

一些比较紧凑的标签地图往往具有和词云非常相似的性质，它们的算法思路对形状词云的生成也有帮助. 显著标签地图[28]对空间上具有一定规律的数据在真实地图上进行展示，以显示每个局部地理实体上哪个数据类别最显著. 算法分为 3 个部分：首先生成一个种子位置的集合，种子从真实的点的位置中采样获得；然后计算各个候选标签的字体大小和标签类别，此时标签之间是有重叠的；最后通过贪婪算法选出一个没有重叠的标签子集. 生成的结果相对 GWC 来说比较稀疏，更倾向于真实地展示地理数据的功能而非追求美观. 类似的工作还有随时间变化的显著标签地图[29]，针对具有地理信息的话题在时序上的兴衰进行连续的展示.

2.2.2　非地理的其他形状词云

WordArt①是一种流行的商业化词云生成工具，可以生成任意形状的词云，其生成的效果比较好，如图 6 所示. WordArt 的缺点在于，为了提高填充率会对个别单词进行放大，在一定程度上破坏了数据真实性. 若强行禁止这种不均匀的放大，其中大单词对形状的填充率就大大降低了. Tagxedo②也是比较有名的形状词语可视化工具，它同样存在数据保真度方面的问题：如果单词无法适应形状，它会自动丢弃一部分单词不进行绘制.

形变词云(morphable word clouds, MWC)[20]通过添加带有 6 个限制条件的刚体动力学系统，将 Wordle 算法生成的单词布局成目标的形状. 限制条件包括不允许单词重叠、单词分布均匀化、单词不能超过形状边界、控制单词的方向、固定某些单词的位置以及保持时序上帧与帧之间的相似性. 但是鉴于力学系统的不稳定性，有一些特定形状的结果会生成失败，因为一个大的单词可能会阻碍其他单词的移动，即使力达到了平衡之后单词也无法达到预期位置. 尤其当生成的初始布局的形状和最终的目标形状差距较大时，往往还需要用户进行额外的手工调整.

图 6　WordArt 工具生成的形状词云

ShapeWordle[18]通过形状感知的阿基米德螺旋线直接让单词排布成目标形状. 传统 Wordle 算法的单词移动方向可分解为一个法向量和一个切线向量，这 2 个向量是互相垂直的. ShapeWordle[18]通过计算形状内对应的距离场引导这 2 个向量的方向，使螺旋线前进的方向沿着形状的轮廓. 该方法可以产生适合任意形状的螺旋线，按照这个螺旋线进行碰撞检测可以将单词直接布局成目标形状. 对于比较复杂的形状，ShapeWordle 用多中心的方法将形状分割为若干个部分，每个部分单独进行形状感知的螺旋线排布，最终效果如图 7 所示. 同时，ShapeWordle 继承了 EdWordle[10]中的可编辑方法，允许用户编辑大单词的大小、位置和旋转角度，编辑完成之后填充小单词.

图 7　ShapeWordle[18]示例

① http://www.wordart.com

② http://www.tagxedo.com

总之，对于在多个设计准则或限制条件之间博弈的算法，数据保真度往往会为美观度而牺牲；直接使用力学模型方便简单，但效果不稳定；而改造传统螺旋线为形状螺旋线的方法为形状词云生成开拓了新的思路.

2.3 可编辑词云

随着计算机和触屏设备的普及，交互式可视化成为了研究热门. 前文提到过的 WordArt 等在线网站允许用户手动选择形状词云的外形轮廓、单词的朝向、词语颜色等. 这些创作工具均能够为一次性生成的词云选定生成参数. 有时用户们在生成词云之后还希望对词云进行编辑修改，而不破坏原有的紧凑性和布局结构. 本节重点介绍对于已经生成的词云中的单词进行再次编辑修改的相关研究. 可编辑词云满足了设计师们定制个性化词云的需求，同时兼顾自动生成的方便快捷和修改控制细节时的高度灵活性.

词云编辑的目标是在保持词云整体紧凑、单词之间没有交叠的前提下，对词云进行局部的重新布局而不损伤词云的美观性. 实现这个目标的难点很容易想象：若将一个单词删掉，则其所在的位置就会产生一个空洞；如果要把一个单词移动到某个位置，则该位置的其他单词就要被移走.

ManiWordle[8]允许用户使用鼠标移动、旋转及删除词云中的单词，界面如图 8 所示. 布局的调整方式遵循大单词更重要的原则，当用户将单词移动到更小单词占有的位置上时，这个更小的单词会被移走，为当前单词腾出空间. 如果移动到更大单词的位置上，这个更大的单词不会被移走. 当人们调整一个单词的大小时，周围发生碰撞的单词会被移动到最近的可塞入的位置. 用户也可以手动单击固定住他们不想移动的单词，避免在编辑过程中被其他单词挤走. 为了消除编辑造成的空白，ManiWordle 最后会对未编辑且未固定的单词重新运行 Wordle 算法. 这种做法能够保证每次编辑并重新排布之后的布局满足美观紧凑的要求，缺点是编辑前后除了被编辑单词和被手动固定的单词之外，其他单词的位置都发生了不可预测的改变，特别是删除单词这一操作，删除前后其他单词位置发生了较大变动.

WordlePlus[9]将 ManiWordle 的交互延伸到了可触控设备上，为之设计了更加自然的交互方式. 当删掉词云中的一个单词时，WordlePlus 重复使用词云边界附近的小单词对词云中的空白进行填充，这种方法同样会造成全局性的不可预测变化.

图 8 Maniwordle[8]的交互界面

EdWordle[10]提出了另一种解决方案：使用刚体动力学系统将每个单词看做一个物理意义上的刚体. 刚体与刚体之间是不可以发生重叠的，发生重叠时会弹开. 对单词刚体施加朝向画布中心的力和保持单词互相之间关系的邻居力，每次编辑之后，这个动力学系统就会重新计算一次达到力学上的平衡，被编辑单词的局部变化被整体的力学动态平衡的系统消化掉，使整体布局稳定保持下来. EdWordle 示例如图 9 所示. 除了允许用户定制个性化词云这一用途之外，其还可以用于优化已经生成的现有其他词云，如让生成的语义词云在保持单词之间相对位置关系的基础上更加紧凑.

图 9 EdWordle[10]编辑前后效果示例

作为一种非常普及的文本可视化工具，词云的定制化、个性化已经成为越来越重要的需求. 如何通过更好的交互方式方便用户进行个性化词云创作，是有待通过实验验证并进一步尝试的问题.

2.4 多文档词云

传统词云的方法都是使用一段文本作为输入，生成一张单独的词云视图. 本节讨论算法输入是多个独立文档的词云，主要分为 2 种类型：第 1 种是对每个文档生成一张单独的词云视图，这种情况需要保持不同文档之间的一致性；第 2 种是用于探索大规模文档集的词云，利用所有文档共同生成的一张词云主视图，引导人们快速地获得关于文档集的见解.

2.4.1 多视图保持一致性的词云

对于每个文档单独生成一张词云视图的情况，基本要求是多个文档中重复出现的单词要排布在各个文档对应词云中相近的位置上. 对于时序多

文档数据，这种视图之间的一致性有利于人们追踪单词发生的变化；对于相似文档类型数据，这有利于用户进行比较．目前这一要求的实现途径主要包括: (1) 多个词云视图联合优化; (2) 带有限制的力学模型; (3) 设计一个应用于所有词云视图中的共同联合布局等．

优化算法是解决这一问题可能途径．Word storms[12]最初被发明时是被用来做文档比较的，其输入是多篇相似文档．算法核心是通过求解优化方程得出单词位置和大小等排布信息．它的缺点是将单词大小也纳入了优化过程中，因此最终展示的单词大小可能并未反映出文本中真实的词频大小．

鉴于力学模型天然的保持邻居一致性特性，很多研究者尝试通过力导向控制单词位置不变．WordSwarm①利用力学系统对多文档词云进行了实现，通过实时计算的力学引擎保证帧与帧之间单词的平滑移动，在数据变化较小的情况下有比较好的效果．由于力作用在物体上引起物体运动需要一定的时间，因此，这种实现方式无法应对帧与帧之间单词变化特别剧烈的情况，往往会因为力来不及作用完全而形成单词之间的重叠，或者造成空白．MWC[20]使用一系列形状约束生成时序词云，为刚体动力学系统添加时序上保持一致性的限制条件；同样，因为依赖力学系统，该方法的稳定性无法得到完全保证．

创建应用于所有词云视图的共同联合布局是目前大多数人采用的思路．保持上下文的动态词云[11]所使用的共同联合布局是它在初始时通过将所有单词多维度缩放生成的布局．该方法在每一帧中构造出图，图的各点都是这个共同联合布局中的点的子集．这种方法的缺点是，如果不同时间步之间单词发生的变化比较剧烈(指不同文档之间单词的消失和出现变化比较大，或者共有单词的权重变化大)，加力之后的结果可能会和预期的一致性差距比较大．最终，将每个时间点的词云与一张折线图关联起来，折线图的横纵坐标分别表示时间和当前时间点对应词云包含的信息量．用户可以通过点击交互逐个地展示时间序列上的词云．

在构建共同联合布局时，Seyfert 等[39]修改了原 Wordle 螺旋线算法中单词排序方法，试图保持整体布局的稳定．单词的颜色和倾斜角度都代表了从上一时间到当前时间单词大小发生的变化．最终结果以动画形式展示，并通过更细粒度的碰撞检测去除单词之间的重叠，计算消耗量较大．

Tag River[30]将标签云与流图相结合，为时序文本数据提供了一个用来完成比较任务的概览．使用相邻的多个狭长多边形区域组成流图，用来展示时间上的整体趋势；当前用户观察的时间段用标签云进行具体的文本展示．用户可以通过交互选择其他感兴趣的时间段，标签云随之以动画的形式变动至对应时间段的内容位置．为了保证变动前后标签布局的一致性，Tag River 使用了装箱算法尽量减小标签占用的总面积空间，为每个单词提供了允许其发生变化的缓冲区．为了提高填充率，有时需要对单词进行缩小，改变原有的数据，否则会产生过于稀疏的布局．

Herold 等[21]提出基于文档比较目的设计的词云，输入的是因用户交互而不断产生的流数据，因而无法进行全局布局优化．该方法考虑了单词的共现关系，将不同时出现的单词进行编组放置在同一位置节省空间．尽管从提供的效果来看最终实现的布局非常稀疏，有时单词与单词之间会出现重叠，这种利用原始数据模式的方法仍然是值得关注的一种思路．

2.4.2　文档集的单一视图词云

利用词云的空间布局可视化大规模的文档集合，有助于用户快速地获得文档集整体信息的概览．

有一些工作通过引入其他可视化元素辅助词云展示文档集合的信息，如图 10 所示 Parallel tag clouds[13]引入了平行坐标系，通过连线展示时序信息；如图 11 所示 SparkClouds[14]引入折线图，在每个标签下面添加反映该标签变化的短线；还有人将展示单词时序变化的直方图引入标签云[15]，当用户选择标签云中的一个单词时，与之一起出现的单词会使用背景高亮显示，附以共现的次数，对应每个单词的直方图中共现的时间段对应的部分被高亮显示．

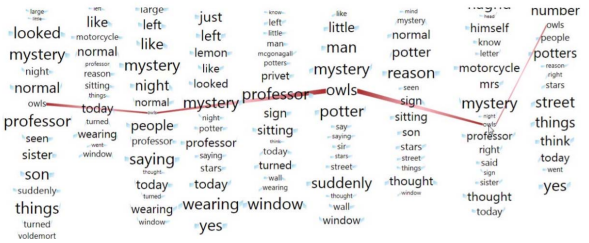

图 10　Parallel tag clouds[13]

① https://github.com/thisIsMikeKane/WordSwarm

图 11 SparkClouds[14]

以词云或类似词云的界面作为主要视图，辅以用户交互，帮助人们探索大规模文本集合. 这种词云中单词之间的邻居关系有时会被赋予语义上的层次和距离含义，如按相似度聚类，以方便人们进行浏览[31].

一些工作将词云与树图结合起来，对文档集分层次地进行展示. NewsMap①与树图相结合用来展示新闻消息. ThemeCrowds[32]层次地概括展示了推特上用户讨论的不同话题，允许用户通过交互查看他们感兴趣的不同层次尺度的话题，主要包括 3 个组成部分: (1) 用户输入搜索词的搜索框; (2) 展示当前话题相关度和尺度的滚动条，横轴代表时间，纵轴代表当前话题的推特的量，颜色代表话题与用户搜索词的相关度; (3) 以树图形式展示的多层标签云，用户可以通过点击树图的各个节点展开更深一层的标签云内容.

研究人员常将词云与流图结合起来. TIARA[33]展示了多个文本主题各自单独随时间的变化情况. 流图的每层都代表了一个从文档集合中提取出的文本主题，x 轴代表时间，y 轴代表文本主题的强度(当前主题的文档数量). 如图 12 所示，每个文本主题提取出其在各个时间点的关键词，并以多个词云的形式排布到流图对应时间点的位置上，允许用户通过交互探索分析文档集合. 例如，用户可以点击关键词查看包含关键词的相关文档来理解关键词与主题之间的关系; 用户在搜索框内输入他们想要查找的关键词，系统将为他们可视化关键词相关的主题; 用户还可以通过交互修改文本分析的结果，甚至人为修正初始生成的主题模型.

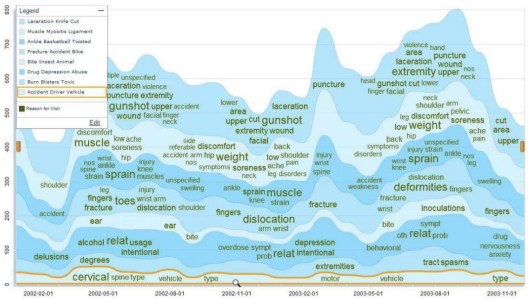

图 12 TIARA[33]

类似的工作还有前文提到过 Tag River[30]，其在结合词云与流图的基础上添加了交互和动画辅助展示主题演变，区别在于 Tag River 在当前视图中仅展示当前时间段的内容，其他时间段的内容会被折叠隐藏.

TextFlow[34]是通过流图展示了文本数据挖掘得到的多主题演变趋势. 流图中的每条流区域代表了一个主题，其宽度与相关文本数量成正比; 主题可能包含产生、融合、分裂和终结这几个演变阶段，对应图中流的起始、合流、分流和结束. 通过该工具可以探索到一些有用的多主题演变信息. 例如，合流和分流情况较多暗示相关主题在对应时间段内发生了较大变动.

RoseRiver[35]是在 TextFlow 基础上的一种渐进式地探索和分析具有层次关系的文本主题的方法，其引入了一种“演变主题树”的结构用来帮助人们理解时序大规模文档集合. 用户可以通过交互修改剪枝，渐进式地探索文本结构.

类似地，CiteRivers[40]通过结合流图和词云的方式展示科学文献. 根据文献之间的相似程度将文献聚类，同类文献布局在同一条流中，并从中抽取出关键词组成词云.

Typograph[36]将词项、短语项、片段和文档 4 个层次的细节集中到一张词云视图中进行展示，并且保持了单词之间的相似度，因而最终视图中文字组成的簇反映了文本的各个主题. 依据“整体概览，缩放及过滤，细节展示”的思想设计交互，让用户能够通过放大操作渐进地可视化并分析文本，逐步展开文档集合的细节.

进一步，TexTonic[37]在 Typograph 的基础上，允许用户通过与词云界面交互影响系统中的数据模型，如图 13 所示. 系统包含 3 个基础的组成部分. (1) 底层的分析和数据处理，针对文档集合自动生成初始的数据模型. 提取出关键词和关键短

① http://newsmap.jp/

语，生成相似度矩阵，使用 *k*-means 方法生成一系列簇，对簇的中心使用主成分分析进行降维，将簇的中心固定在二维空间的一个位置上，对每个簇中剩下的单词施加力导向模型确定坐标位置．(2) 可视化界面，实时更新用户交互造成的影响．用户交互造成的变化会以动画的形式显示出来．同一个语义簇使用相同的背景上色．可以悬停鼠标查看文本层次的关系．(3) 用户引导的语义交互．用户可以放缩、固定、移动和删除单词，系统实时更新单词的权重和邻居关系．TexTonic 不像静态词云那样完全不允许单词重叠，用户可以通过交互自由调整单词之间的距离．

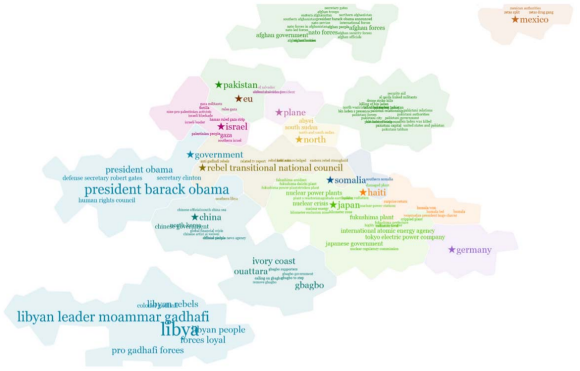

图 13　TexTonic[37]

3　词云的实验评价

对词云的评价通常围绕量化评价、用户实验和案例分析 3 个方面进行．量化评价通过定义度量公式对不同的词云的优劣进行比较；用户实验用于探索用户对词云的感知，词云的设计空间以及用户主观偏好等等；案例分析就是对于自己的或前人的词云生成方法产出的实例进行描述分析，以展示其优劣．就案例分析的方法而言，词云与可视化领域其他方向的工作差异不大，因此本节仅重点讨论前 2 个方面．

3.1　量化评价

文献[41]量化评估了 Wordle，CPD，SC 以及其他 3 种作者自己提出的语义词云方法在保持单词邻居关系、语义距离、布局紧凑程度、均匀度、长宽比以及运行效率这 6 个方面的表现．文献[41]定义的这 6 种度量标准，为之后的语义词云和其他词云方法都提供了良好的评价参考准则[10,18]．

3.2　用户实验

目前大部分研究者都认为词云的娱乐性大于功能性，适用于演示、展览的场景，而不适用于精确的数据分析[2,10]．词云在其优雅外观之下也隐含了一些不容忽视的缺点，如 Wordle 算法布局破坏了自然的阅读顺序、使用字体大小表达量化的词频信息并非最优的视觉通道、人们对单词大小判断的天然偏差等问题．前人致力于通过用户实验了解人们对词云的感知，探讨在词云引人注目的外观基础上如何增强其功能性，扬长避短，提高其表达能力．

Rivadeneira 等[5]做过关于标签云的用户实验，认为标签云可以支持的任务有搜索目标单词、浏览(无固定目标的阅读)、文本印象形成(获取文本主题)、识别主题．他们总结的这些词云任务也被后来的一些研究人员所沿用或引申[6-7]．他们设计了 2 个实验，研究了标签云的不同文字特征和单词布局方式对人们完成特定任务的影响．研究人员发现，大单词比小单词更容易被回想起来，第 1 象限中的内容比其他象限的内容更容易被回想起来．人们对词云不同位置感知程度不同，这一点后来也被其他实验研究反复印证[6,42-44]．研究发现，单词布局方式对主题匹配的任务几乎没有影响，但是对主题发现任务有影响，按词频排序的单列表布局在这一任务中是最高效的，Wordle 算法布局其次．

之后，文献[7]更为详细地分解了标签云中的各个可视特征，通过让用户选择标签云中最重要的单词研究哪种特征最吸引用户注意力．其列出了 9 种可视特征，其中字体大小和字体权重(粗细体)最能吸引用户注意力，其次是透明度和颜色；其他因素(如像素个数、标签宽度、标签占用面积等)不被用户看重．

另有研究[42]验证了按照字母表顺序排序可以帮助用户快速地找到信息，并且认为用户在观察标签云时的行为是浏览而非传统阅读．一些研究人员将眼球追踪技术运用到了标签云的用户实验中[43-44]，从视觉关注焦点的角度进一步验证了前人得出的结论．

Alexander 等[45]研究了将字体大小作为数据编码方式对人造成的偏差．结论是用户对文字单词大小的感知是有偏差的，人们往往会将字体大小与单词的长度、高度和宽度合并理解．实验探究的问题就是单词外形上的特征多大程度会影响对单词大小的感知．实验任务是让用户在众多大小不同的单词构成的词云中，针对用颜色标记出的 2 个单词，选择出哪个更大，将准确率作为需要度量的量．实验结论是单词的长度(字符个数)、高度和实

际宽度都会影响用户对于单词大小的判断，其中单词长度的影响并非字符个数本身的问题，本质上是因为单词长度影响了单词的实际宽度．另外值得注意的是，Alexander 等在实验讨论的部分中特别提到，尽管有高低波动，总体来说用户选中正确单词的准确率其实还是相对比较高的，他们认为用户对使用词云进行大小比较这一任务的完成比前人认为的要更好．

Felix 等[6]系统地探究了词云的设计空间．实验涉及的任务包含底层的数值大小判断和关键词搜索，以及 2 个高层任务：主题匹配和主题发现．他们将词云的设计空间划分为布局策略和大小编码策略 2 个方面．实验涉及的布局策略包含成行排布、成列排布和 Wordle 排布 3 种；实验涉及的大小编码策略从字体性质和附加图形 2 个角度出发，字体性质包含字体大小和字体颜色，附加图形考虑了添加柱或者圆辅助人们获取值．(1) 在大小判断的实验中，用户需要给出词云中 2 个高亮单词相对大小的确切数量关系．实验结果是添加附加图形的编码方式表现较好，布局策略中成列排布的表现较好．(2) 在搜索关键词的实验中，实验人员记录用户找到一个给定单词需要花费的时间．他们发现字体大小和颜色的编码方式明显更优，布局策略中 Wordle 布局和成列布局的表现都很好．同时他们还印证了前人的结论：如果对词云划分象限，那么从上到下、从左到右人们的表现会逐渐变差．(3) 第 3 个主题匹配和第 4 个主题发现任务，各种不同的设计方案之间没有明显差异．Felix 等的结论是，随着任务难度的提升，各种情况之间的差异变得几乎消弭．综合所有实验结果，没有哪种设计方案是能够完胜其他方法的．换言之，用户的表现总是和任务相关的．文献[6]仅从任务完成效率的方面进行分析，没有考虑词云的美观度．

文献[17]中的用户实验研究主要针对语义词云．作者设计的任务类似于一种称为 Taboo 的游戏：提供 5 个反映同一个文本主题的单词作为线索，让用户猜测这 5 个单词暗示的主题是什么，具体示例如图 14 所示．实验涉及 Wordle 布局、成列布局和 SC 布局等，考虑了字体大小有无差异及字体是否上色等多个变量．结果证明人们在成列布局、字体大小全部相同和有颜色对类别进行区分的情况下表现更好．从用户的主观反馈上看，人们普遍认为成列布局的功能性更强，而 Wordle 布局的美观性更佳．研究人员做了眼球追踪的实验，发现成列排布表现更优的原因是它们将单词在空间上聚类排布了，让观察者能够每次看一个类别．综合所有实验的结果，结论是对于语义词云，最好按照语义组织单词，将它们在视觉上进行分隔和上色．

图 14 文献[17]中实验使用的词云示例

可以看到，不管是对于语义词云还是普通词云，从用户实验的结果来看，似乎行列排列的布局从功能性上讲表现更优．然而，因为词云本身在实际运用中并不会被苛求支持高效完成数据分析任务，所以更美观的 Wordle 布局受欢迎程度明显更高．大多数人使用词云的首要目标仍然是希望它能起到吸引注意力的作用，进而使观众从中发现他们感兴趣的内容．不过，尽管在现实生活中人们并不会使用词云完成上述研究中的一些底层感知任务，但是高层感知任务往往依赖于底层感知去完成，仍然可以将研究人员们设计的任务和实验看做是了解人们对词云感知的一个重要参考．

4 结 语

本文总结了词云在近 20 年的研究工作．首先介绍了词云的设计空间，包括视觉编码、布局方法和交互方式 3 个方面；随后按照类别介绍了目前流行的 4 种词云的研究方向：(1) 语义词云通过将单词的语义联系转化为二维平面上的空间联系，增强词云的语义表达能力；(2) 形状词云通过为词云增加形状限制产生更加优美的外形和更加直观的含义表达；(3) 可编辑词云作为最近流行的概念，为词云设计者们提供了更高的自由度，使个人定制词云成为可能；(4) 多文档词云呈现了文档集合更加丰富的含有变化的信息．研究者们在开发这些丰富的词云算法的同时，也在不断明确词云的设计空间；有一些词云除了使用文字大小作为编码方式之外，还选择了其他通道，如颜色、透明度，或者可以添加附加符号，如折线，也会对词云感知

产生影响；此外，也有研究探索词云可能的评价方式和人们对词云的感知能力．在词云被广泛使用的今天，如同文献[6]所言，词云使用者们应关注词云适用的场合和任务，恰当地使用而非滥用词云．

通过总结发现，在词云领域仍然有很多可以探索的研究方向．(1) 目前大多数语义词云的工作仍然依赖自然语言处理的方法表示单词，并使用各种不同的降维算法将单词投影到二维空间，按照这种思路完成的词云布局可能会出现局部难以解释的单词邻居关系，如原本语义上关联性不大的单词恰好被摆在了靠近的位置．未来研究可能需要更为精细的局部布局方法取代难以控制的降维布局方法．(2) 对于形状词云，Wang 等[18]对螺旋线的改进算法已经能够较好地模拟出轮廓外形．对于某些特定的图形，未来可以通过进一步使用单词模拟形状细节达到更好的美观效果，如填充鸽子形状的词云时用单词模拟单词羽毛的方向．总之，对于形状词云，更高的美观性是研究者们的不断追求的目标．(3) 对于多文档词云，目前可研究的范围更为广泛．对于相似文档数据，如何利用词云作为可视化主要界面或辅助工具的情况下，充分展示共同点和不同点，突出论点和主题，都是有待探索的问题．这不仅仅依赖自然语言处理技术的提升，更需要有优秀的展示方法．对于时序文本数据，目前已有的方法难以兼顾稳定性和紧凑性．尤其在当前，不断到来的流式数据给词云可视化带来了新的挑战．(4) 目前几乎所有的研究工作都是基于英文文本的，对中文词云的感知方面目前也没有研究．中文单字由笔画构成，字形上也与英文单词的字母不同，在用单词进行高密度的空间填充时会产生与英文单词不同的效果；中文需要多字组合构成词，构造语义词云时可考虑重复字的利用，生成更符合中国人识别词语习惯的词云．

参考文献(References):

[1] Bausch P, Bumgardner J, Fake C. Flickr hacks: tips & tools for sharing photos online[M]. Boston: O'Reilly Media, Inc, 2006

[2] Viegas F B, Wattenberg M, Feinberg J. Participatory visualization with Wordle[J]. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6): 1137-1144

[3] Feiberg J. Wordle[M] //Steele J, Iliinsky N. Beautiful visualization: looking at data through the eyes of experts, vol 3. 1st ed. Sebastopol : O'Reilly Media, Inc, 2010: 37-58

[4] McNaught C, Lam P. Using Wordle as a supplementary research tool[J]. The Qualitative Report, 2010, 15(3): 630-643

[5] Rivadeneira A W, Gruen D M, Muller M J, *et al*. Getting our head in the clouds: toward evaluation studies of tagclouds[C] // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: Association for Computing Machinery, 2007: 995-998

[6] Felix C, Franconeri S, Bertini E. Taking word clouds apart: an empirical investigation of the design space for keyword summaries[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 657-666

[7] Bateman S, Gutwin C, Nacenta M. Seeing things in the clouds: the effect of visual features on tag cloud selections[C] // Proceedings of the 19th ACM Conference on Hypertext and Hypermedia. New York: Association for Computing Machinery, 2008: 193-202

[8] Koh K, Lee B, Kim B, *et al*. ManiWordle: providing flexible control over wordle[J]. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 1190-1197

[9] Jo J, Lee B, Seo J. WordlePlus: expanding Wordle's use through natural interaction and animation[J]. IEEE Computer Graphics and Applications, 2015, 35(6): 20-28

[10] Wang Y H, Chu X W, Bao C, *et al*. EdWordle: consistency-preserving word cloud editing[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 647-656

[11] Cui W W, Wu Y C, Liu S X, *et al*. Context-preserving dynamic word cloud visualization[J]. IEEE Computer Graphics and Applications, 2010, 30(6): 42-53

[12] Castella Q, Sutton C. Word storms: multiples of word clouds for visual comparison of documents[C] //Proceedings of the 23rd International Conference on World Wide Web. New York: Association for Computing Machinery, 2014: 665-676

[13] Collins C, Viegas F B, Wattenberg M. Parallel tag clouds to explore and analyze faceted text corpora[C] //Proceedings of the IEEE Symposium on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2009: 91-98

[14] Lee B, Riche N H, Karlson A K, *et al*. SparkClouds: visualizing trends in tag clouds[J]. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 1182-1189

[15] Lohmann S, Burch M, Schmauder H, *et al*. Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds[C] //Proceedings of the International Working Conference on Advanced Visual Interfaces. New York: Association for Computing Machinery, 2012: 753-756

[16] Nguyen D Q, Tominski C, Schumann H, *et al*. Visualizing tags with spatiotemporal references[C] //Proceedings of the 15th International Conference on Information Visualisation. Los Alamitos: IEEE Computer Society Press, 2011: 32-39

[17] Hearst M A, Pedersen E, Patil L, *et al*. An evaluation of semantically grouped word cloud designs[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(9): 2748-2761

[18] Wang Y H, Chu X W, Zhang K Y, *et al*. ShapeWordle: tailoring Wordles using shape-aware archimedean spirals[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 26(1): 991-1000

[19] Witkin A. Physically based modeling: principles and practice constrained dynamics[J]. Computer Graphics, 1997, 11-21

[20] Chi M T, Lin S S, Chen S Y, *et al*. Morphable word clouds for time-varying text data visualization[J]. IEEE Transactions on Visualization and Computer Graphics, 2015, 21(12): 1415-1426

[21] Herold E, Pöckelmann M, Berg C, *et al*. Stable word-clouds for visualising text-changes over time[C] //Proceedings of International Conference on Theory and Practice of Digital Libraries. Heidelberg: Springer, 2019: 224-237

[22] Wu Y C, Provan T, Wei F R, *et al*. Semantic-preserving word clouds by seam carving[J]. Computer Graphics Forum, 2011, 30(3): 741-750

[23] Paulovich F V, Toledo F M B, Telles G P, *et al*. Semantic wordification of document collections[J]. Computer Graphics Forum, 2012, 31(3pt3): 1145-1153

[24] Schubert E, Spitz A, Weiler M, *et al*. Semantic word clouds with background corpus normalization and t-distributed stochastic neighbor embedding[OL]. [2020-11-24]. https://arxiv.org/pdf/1708.03569.pdf

[25] Wang J, Zhao J, Guo S, *et al*. ReCloud: semantics-based word cloud visualization of user reviews[C] //Proceedings of Graphics Interface. Toronto: Canadian Information Processing Society, 2014: 151-158

[26] Xu J, Tao Y B, Lin H. Semantic word cloud generation based on word embeddings[C] //Proceedings of the IEEE Pacific Visualization Symposium. Los Alamitos: IEEE Computer Society Press, 2016: 239-243

[27] Buchin K, Creemers D, Lazzarotto A, *et al*. Geo word clouds[C] //Proceedings of the IEEE Pacific Visualization Symposium Los Alamitos: IEEE Computer Society Press, 2016: 144-151

[28] Reckziegel M, Cheema M F, Scheuermann G, *et al*. Predominance tag maps[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(6): 1893-1904

[29] Reckziegel M, Jänicke S. Time varying predominance tag maps[C] //Proceedings of the IEEE Visualization Conference. Los Alamitos: IEEE Computer Society Press, 2019: 231-235

[30] Forbes A G, Alper B, Höllerer T, *et al*. Interactive folksonomic analytics with the tag river visualization[C] //Proceedings of the IEEE Workshop on Interactive Visual Text Analytics. Los Alamitos: IEEE Computer Society Press, 2011: 3

[31] Hassan-Montero Y, Herrero-Solana V. Improving tag-clouds as visual information retrieval interfaces[C] //Proceedings of International Conference on Multidisciplinary Information Sciences and Technologies. Catonsville: Institute for Operations Research and the Management Sciences, 2006: 25-28

[32] Archambault D, Greene D, Cunningham P, *et al*. ThemeCrowds: multiresolution summaries of twitter usage[C] //Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents. New York: Association for Computing Machinery, 2011: 77-84

[33] Wei F R, Liu S X, Song Y Q, *et al*. TIARA: a visual exploratory text analytic system[C] //Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2010: 153-162

[34] Cui W W, Liu S X, Tan L, *et al*. TextFlow: towards better understanding of evolving topics in text[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2412-2421

[35] 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

[36] Endert A, Burtner R, Cramer N, *et al*. Typograph: multiscale spatial exploration of text documents[C] //Proceedings of the IEEE International Conference on Big Data. Los Alamitos: IEEE Computer Society Press, 2013: 17-24

[37] Paul C L, Chang J, Endert A, *et al*. TexTonic: interactive visualization for exploration and discovery of very large text collections[J]. Information Visualization, 2019, 18(3): 339-356

[38] Kruskal J B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis[J]. Psychometrika, 1964, 29(1): 1-27

[39] Seyfert M, Viola I. Dynamic word clouds[C] //Proceedings of the 33rd Spring Conference on Computer Graphics. New York: Association for Computing Machinery, 2017: 1-8

[40] Heimerl F, Han Q, Koch S, *et al*. CiteRivers: visual analytics of citation patterns[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 190-199

[41] Barth L, Kobourov S G, Pupyrev S. Experimental comparison of semantic word clouds[C] //Proceedings of International Symposium on Experimental Algorithms. Heidelberg: Springer, 2014: 247-258

[42] Halvey M J, Keane M T. An assessment of tag presentation techniques[C] //Proceedings of the 16th International Conference on World Wide Web. New York: Association for Computing Machinery, 2007: 1313-1314

[43] Lohmann S, Ziegler J, Tetzlaff L. Comparison of tag cloud layouts: task-related performance and visual exploration[C] // Proceedings of IFIP Conference on Human-Computer Interaction. Heidelberg: Springer, 2009: 392-404

[44] Schrammel J, Deutsch S, Tscheligi M. Visual search strategies of tag clouds-results from an eyetracking study[C] // Proceedings of IFIP Conference on Human-Computer Interaction. Heidelberg: Springer, 2009: 819-831

[45] Alexander E, Chang C C, Shimabukuro M, *et al*. Perceptual biases in font size as a data encoding[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(8): 2397-2410