文本可视化研究综述

唐家渝, 刘知远, 孙茂松
(智能技术与系统国家重点实验室,清华信息科学与技术国家实验室(筹)清华大学计算机科学与技术系 北京 100084)
(tjy430@gmail.com)

摘 要: 随着海量文本的涌现,信息超载和数据过剩等问题促使了文本可视化技术的出现. 文本可视化技术综合了文本分析、数据挖掘、数据可视化、计算机图形学、人机交互、认知科学等学科的理论和方法,为人们提供了一种理解复杂文本的内容、结构和内在规律等信息的有效手段. 文中首先阐述了文本可视化的概念和重要性,然后按照不同可视化对象类型综述了文本可视化的研究现状,并介绍了典型的文本可视化方法与方案;最后,对文本可视化的未来研究方向进行了展望.

关键词: 信息可视化;可视分析;文本分析;信息抽取;人机交互界面
中图法分类号: TP391

A Survey of Text Visualization

Tang Jiayu, Liu Zhiyuan, and Sun Maosong
(*State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract: With the emergence of massive texts, information overload and data redundancy raise great challenges for information processing. To address these issues, text visualization has been proposed for understanding the content, structure and patterns hidden behind complicated textual information. Text visualization integrates several techniques including text analysis, data mining, data visualization, computer graphics, human computer interaction, cognitive science and so on. In this paper, we first introduce the concepts of text visualization. Afterwards, we present the research achievements according to different visualization objects, and introduce typical visualization methods and schemes. As a conclusion, we give an outlook to future research directions of text visualization.

Key words: information visualization, visual analytics, text analysis, information extraction, human-computer interaction interface

1 文本可视化的定义

随着信息技术的快速发展,海量信息不断涌现,使得人们对其处理和理解的难度日益增大. 传统的文本分析技术虽已在一定程度上实现了从大数据中挖掘出重要信息,但是这些挖掘出的信息通常仍然无法满足人们利用浏览及筛选等方式对其进行合理的分析、理解和应用. 面对这种挑战,文本可视化技术应运而生,它将文本中复杂的或者难以通过文字表达的内容和规律以视觉符号的形式表达出来,同时向人们提供与视觉信息进行快速交互的功能,使人们能够利用与生俱来的视觉感知的并行化处理能力快速获取大数据中所蕴含的关键信息. 文本可视化

收稿日期:2012-12-31. **基金项目**:国家自然科学基金重点项目(61133012).**唐家渝**(1987—),男,硕士研究生,论文通讯作者,主要研究方向为信息可视化;**刘知远**(1984—),男,博士后研究人员,CCF 会员,主要研究方向为关键词抽取、社会标签分析和社会计算;**孙茂松**(1962—),男,博士,教授,博士生导师,CCF 理事,主要研究方向为自然语言处理、信息检索和社会计算.

综合了文本分析、数据挖掘、数据可视化、计算机图形学、人机交互、认知科学等学科的理论和方法，为人们提供了一种理解海量复杂文本的内容、结构和内在规律等信息的有效手段.

2 文本可视化的基本方法

文本可视化涵盖了信息收集、数据预处理、知识表示、视觉呈现和交互等过程. 其中，数据挖掘和自然语言处理等技术充分发挥计算机的自动处理能力，将无结构的文本信息自动转换为可视的有结构信息；而可视化呈现使人类视觉认知、关联、推理能力得到充分发挥. 因此，文本可视化有效地综合了机器智能和人类智能，为人们更好地理解文本和发现知识提供了新的有效途径.

图 1 展示了人们利用文本可视化系统对文本进行分析和理解的基本过程. 总的来说，文本可视化系统主要包括 3 个部分：1)产生可视化所需数据的文本分析过程；2)可视化呈现，即包含文档、事件、关系或时间等文本信息的低维信息图(通常是 2D 或 3D 图)；3)用户与信息图的交互.

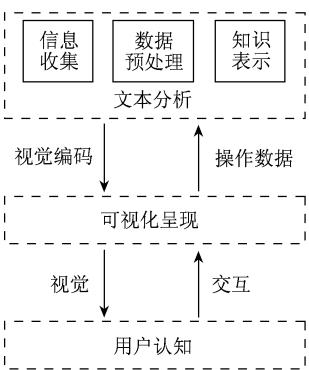

图 1 文本可视化的基本框架

2.1 文本分析

文本可视化依赖于自然语言处理，因此词袋模型、命名实体识别、关键词抽取、主题分析、情感分析等是较常用的文本分析技术.

文本分析的过程主要包括特征提取，通过分词、抽取、归一化等操作提取出文本词汇级的内容；利用特征构建向量空间模型(vector space model, VSM)[1]并进行降维，以便将其呈现在低维空间，或者利用主题模型处理特征；最终以灵活有效的形式表示这些处理过的数据，以便进行可视化呈现和交互.

2.2 可视化呈现

信息图中，文本内容的视觉编码主要涉及尺寸、颜色、形状、方位、纹理等，文本间关系的视觉编码主要涉及网络图、维恩图、树状图、坐标轴等. 文本可视化的一个重要任务是选择合适的视觉编码呈现文本信息的各种特征；例如词语的频度通常由字体的大小表示，不同的命名实体类别用颜色加以区分.

如何快速创建符合人们先验认知的视觉呈现一直是可视化研究者关心的问题[2]，对于视觉编码有效性的研究与认知科学息息相关.

2.3 交 互

为了使用户能够通过可视化有效地发现文本信息的特征和规律，通常会根据使用场景为系统设置一定程度的交互功能. 文本可视化中，主要应用到的交互方式有高亮(highlighting)、缩放(zooming)、动态转换(animated transitions)[3]、关联更新(brushing and linking)、焦点加上下文(focus+context)[4]等.

3 文本可视化的研究现状

文本可视化的研究主要包括基于文本内容的可视化、基于文本关系的可视化、基于多层面(multi-faceted)信息的可视化.

3.1 基于文本内容的可视化

基于文本内容的可视化主要关注的是如何快速获取文本内容的重点，主要可以分为基于词频的可视化和基于词汇分布的可视化. 基于文本内容的可视化可以应用于单个文本，也适用于较大的文本集. 通过这些基本统计结果的可视化呈现，能使用户快速地了解文本的大体内容，这对于进一步的分析具有重要的向导意义.

3.1.1 基于词频的文本可视化

当面对海量文本时，人们需要对每个文本或者整个文本集合的主要内容进行快速浏览，因此需要基于词频的文本可视化. 最常用的文本可视化的思路是将文本看作一个词汇的集合(词袋模型)，利用词频信息来呈现文本特征. 其中，经常被采用的词频计算方法是 TFIDF[5]，最典型的可视化形式是“标签云(tag cloud)[6]”. 标签云将关键词按照一定顺序和规律排列，如频度递减、字母顺序等，并以文字的大小代表词语的重要性. 最初的标签云大多都采用将文字一行一行地水平排列的方式，后来渐渐遵循更加美观复杂的布局规则，图 2 所示[7] Wordle① 便是其中最被广泛采用的代表之一. 在 Wordle 中，

① http://www.wordle.net/

词语的布局遵循了严格的条件，使得文字间的空隙得以充分地利用、可视化结果更加美观. 自 Wordle 出现就被广泛应用于报纸、杂志等传统媒体，以及互联网，甚至 T 恤等实物中[7]. 在 Wordle 的基础上，图 3 所示 Tagxedo① 遵循了更为美观和复杂的布局，允许用户选择不同的文字轮廓甚至自定义轮廓. ManiWordle[8] 在 Wordle 的基础上提供了对标签云进行高效编辑的功能，使用户能够得心应手地定制文本可视化呈现形式.

图 2　Wordle

图 3　Tagxedo

然而，由于标签云只是对文本中高频词汇的简单罗列，无法提供连贯的上下文信息. 而 Document Card[9] 在克服这一问题上做出了尝试，它通过自动提取重要文字和图片，将文本信息综合到一系列信息连续的卡片上，使用户能够快速地了解文本的关键信息. 另外，标签云也经常作为辅助的呈现方式出现在一些可视化方案中.

3.1.2　基于词汇分布的文本可视化

基于词汇分布的文本可视化反映了词汇在文本中的分布情况，它主要应用于查询任务. 通常，它将去掉停用词后的词汇建立成索引，图形化地展示用户输入的查询词在文本中的分布情况. TileBars[10] 将文献按段落、章节等划分为文本单元，通过矩形条的灰度显示每个查询词在文献中的分布，使得用户能够通过观察查询词组在相同的文本单元里的同现情况，快速了解文本内容与查询意图的相关度.

3.2　基于文本关系的可视化

基于文本关系的可视化研究文本内外关系，帮助人们理解文本内容和发现规律；常用的可视化形式有树状图和节点连接的网络图.

3.2.1　基于文本内在关系的可视化

基于文本内在关系的可视化主要关注文本的内部结构和语义关系等. Contexter[11] 利用网络图呈现了新闻中的命名实体在同一文本中的同现关系；Word Tree[12] 结合后缀树（suffix tree）的思想，以图 4 所示树状结构呈现了查询词的上下文关系. NETSPEAK[13] 以类似的形式展现了文本集中常见的上下文结构，帮助用户在写作时选择合适的词语.

图 4　Word Tree

如图 5 所示，Phrase Net[14] 从语义层面分析并呈现命名实体在文本内的多种关系，如从属关系、并列关系等. 语义的层次结构也是常见的可视化内容；FanLens[15] 以径向的空间填充方式呈现了命名实体的层次关系；图 6 所示 DocuBurst[16] 以类似的形式让词语通过 Wordnet[17] 中的上下位关系以放射状径向排列，其中字体大小表示词语在文档中的频度.

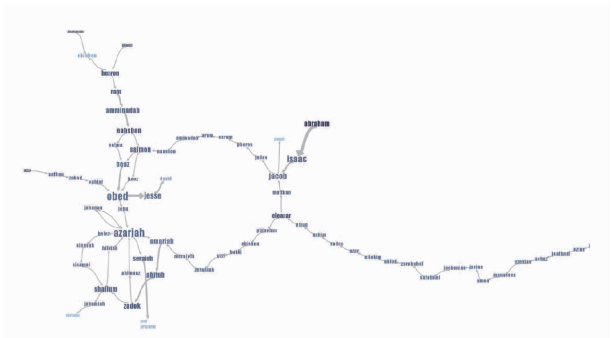

图 5　Phrase Net

① http://www.tagxedo.com/

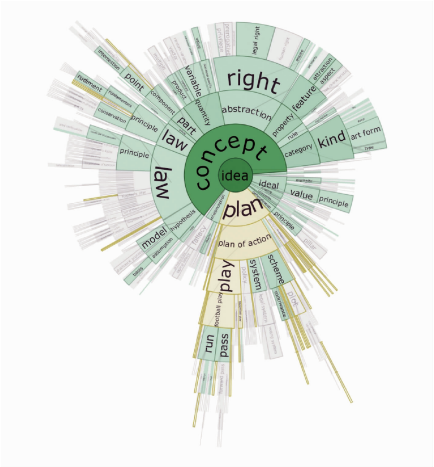

图 6 DocuBurst

3.2.2 基于文本外在关系的可视化

外在关系可视化的内容包括文本间的引用关系、网页的超链接关系等直接关系，以及主题相似性等潜在关系.

许多对于文本间直接关系的可视化研究都以网络图作为呈现形式. 例如，在对文本集的引用关系的可视化中[18-19]，利用网络节点代表文本，用有向的线段表示引用关系. Zhang 等[20]为便于人们进行领域分析，以 FP-tree[21]的结构展现了图 7 所示文献的共引关系[20]，与 CiteSpace[18]这类利用传统网络图的可视化方案相比，它能够呈现出相关文献聚类中更为细致的信息.

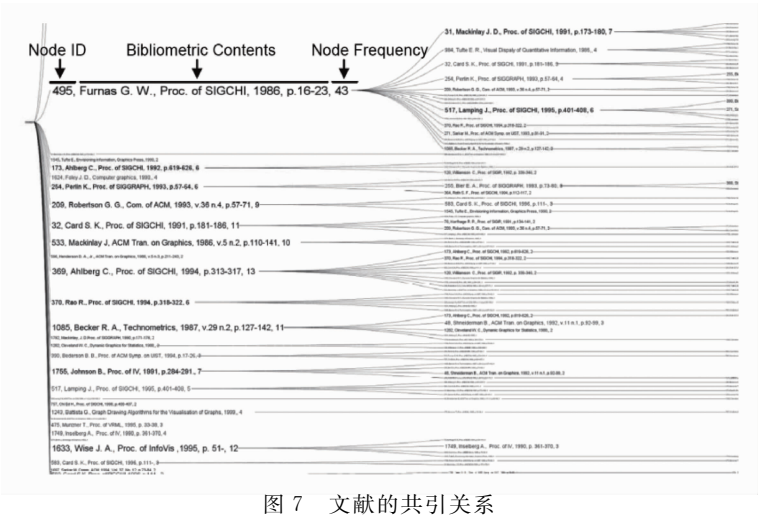

图 7 文献的共引关系

大规模文本集合的主题关系是外在关系更为常见的可视化场景，一般主要基于聚类算法呈现主题分布，并展示与特定主题相关的关键词，多应用于信息检索、主题检测、话题演变等方面.

在前面提到的标签云的生成过程中，文字几乎都是随机排列的，不能保证语义或上下文相关的文字能够按照某种规律排列. Hassan-Montero 等[22]在最传统的按行排列的标签云的基础上，以 Jaccard 系数衡量标签的同现关系来作为聚类依据，同一行的文字表示同一类别，相邻行的类别表示意义相近.

文本主题关系分析除了以上基于统计的方法外，更为常见的是特征降维技术. 该技术首先用高维

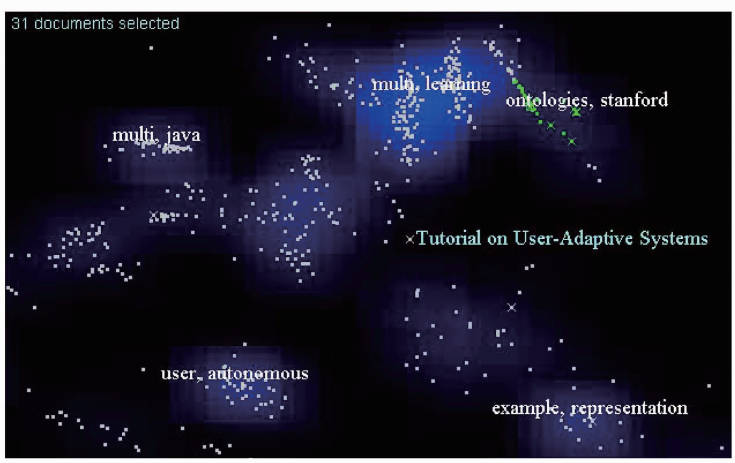

图 8 Galaxy

的 VSM 表示复杂的文本，然后通过投影的方式对文本特征向量进行降维处理，使信息能够在 2D 或 3D 的空间里进行可视化呈现. 常用降维方法包括基于奇异值分解(singular value decomposition，SVD)的潜在语义索引(latent semantic indexing，LSI)[23]、主成分分析(principal components analysis，PCA)[24]、对应分析(correspondence analysis，CA)[25]、多维尺度分析(multidimensional scaling，MDS)[26]以及基于人工神经网络的自组织映射图网络(self-organizing map，SOM)[27].

在基于特征降维技术的可视化中，图 8 所示文本地图[28]是广泛应用的形式[29-35]. ProjCloud[36]进行特征降维时，利用 k-means 算法[37]对文本进行聚类，并结合标签云呈现了文本集合的相似关系和相似文本集合的关键词.

然而，由于降维过程存在信息丢失，导致基于特征降维的可视化也存在一些问题，例如缺乏可扩展性、图形通常过于复杂、文字标签缺乏可读性等[38]. 因此，研究者进而考虑分层次进行文本信息可视化的方案. TreeMap[39]巧妙地使用嵌套的长方形来表示不同层次，以长方形的方向表示不同层次的变换，并以长方形的大小来表示节点的重要性. 图 9 所示 NewsMap① 利用了 TreeMap 的形式展示新闻文本，Map of the Market[40]则呈现了股票市场的概览②. ThemeCrowds[41]将 TreeMap 和标签云结合，同时展现了主题的层次关系和主题的关键词.

图 9　NewsMap

还有一些研究通过生成层次聚类树，借助缩放等交互手段实现了聚类信息的分层级展示[42-44]；其中，InfoSky[42]在布局时使用了力定向布局(force-directed placement，FDP)[45]这种常见的平面可视化技术.

3.3　基于多层面信息的可视化

基于多层面信息的文本可视化主要研究如何结合信息的多个方面帮助用户从更深层次理解文本数据，发现其内在规律；其中，包含时间信息的文本可视化近年来受到越来越多的关注.

3.3.1　基于时间与其他信息结合的可视化

在新闻、博客、邮件、论文等几乎所有文本中，时间都是其重要的属性. 时间信息提供了关于文本内容变化、数据规律等方面的重要信息，因此一直以来是信息可视化中的重要元素[46].

包含时间信息的可视化中，最直接和主要的方式是引入时间轴，并将信息按照时间顺序线性排列. TimeMines[47]，LifeLines[48]和 LifeFlow[49]等均通过将事件从左到右显示在时间轴上，为人们进行基于时间的事件序列分析提供了便捷的途径.

有很多研究试图将标签云与时间相结合，其中，图 10 所示 SparkClouds[50]在标签云的每个词语下方引入折线图，以表示每个词语随着时间的使用频度变化. 通过对标签云上的词语标记不同的颜色和图形也是常用的方式[51-52]，如图 11 所示，Cui 等[53]除了为标签云中的每个词语标记不同的颜色以表示它们随时间的变化情况外，还将每个时间点的标签云与一个折线图相关联，一个标签云本身包含的信息越多，它在折线图上的值就越大.

图 10　SparkClouds

叠式图(stacked graph)是非常常用的可视化形式，Byron 等[54]对美观易读的叠式图的设计与实现技术进行了讨论. 在叠式图中，每层代表一个事物，以不同颜色加以区分，从左到右呈现事物在时间上的变化[55-56]. ThemeRiver[57]利用河流这一隐喻，将时间看作从左到右延续的河流，将文本数据按主题进行分割，堆叠成一张美观的叠式图；其中，每一条彩色线条代表不同主题，每个主题用一个主题词标注，

① http://newsmap.jp/
② http://www.smartmoney.com/map-of-the-market/

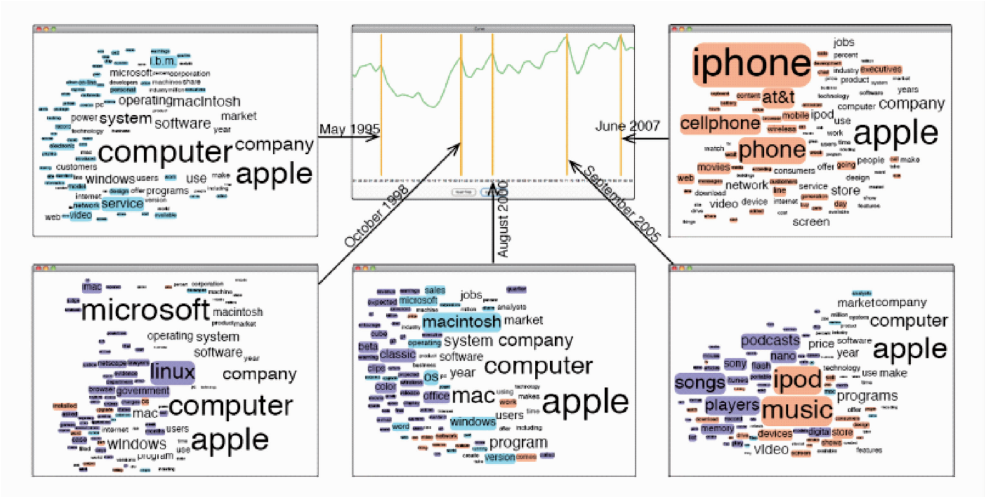

图 11 标签云和折线图的结合

线条的粗细代表主题的频度. ThemeRiver 提出之后产生了很多扩展工作,比较有代表性的有 MemeTracker[58],NewsRiver[59] 和 Visual Backchannel[60],用于跟踪博客、新闻、Twitter 事件的变化.

然而,如图 12 所示,ThemeRiver[57] 由于做了平滑和堆叠处理,较为细节的信息如属性的绝对数值难以识别,因此如何使 ThemeRiver 更加美观和可靠一直是可视化研究的热点. 例如,TIARA[61] 结合了标签云的可视化形式,通过主题分析技术(latent dirichlet allocation, LDA)[62] 抽取文本主题将其展现在 ThemeRiver 中,并将每个主题下的关键词显示在每条线条中,用于展现该主题的详细信息. 此外,Tag River[63] 也利用了河流这一隐喻与标签云结合的方案. 如图 13 所示 TextFlow[64] 以河流的隐喻连贯、细致地展现了主题随着时间的变化,如主题的产生、分解、合并;同时标注了主题变化的关键点和相应时间点的关键词. 利用河流这一隐喻的可视化还有 EventRiver[65].

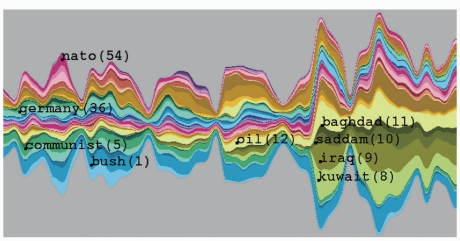

图 12 ThemeRiver

图 14 所示 History Flow[66] 主要研究文档内容是如何随着时间变化的,并对协同共建的百科全书式网站 Wikipedia① 的页面动态性进行了可视化.

螺旋图在包含时间信息的文本可视化中也是较为常见的形式,它能够较好地展现文本数据中的周期性规律[67],同时,多层次的螺旋图还能方便地就不同数据集的周期性规律进行对比[68].

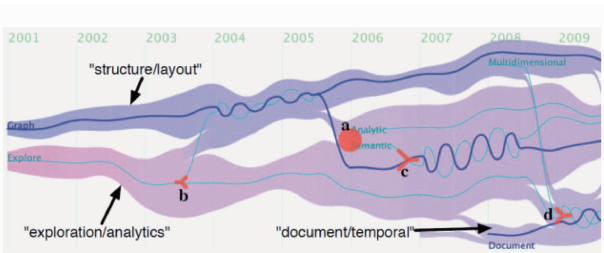

图 13 TextFlow

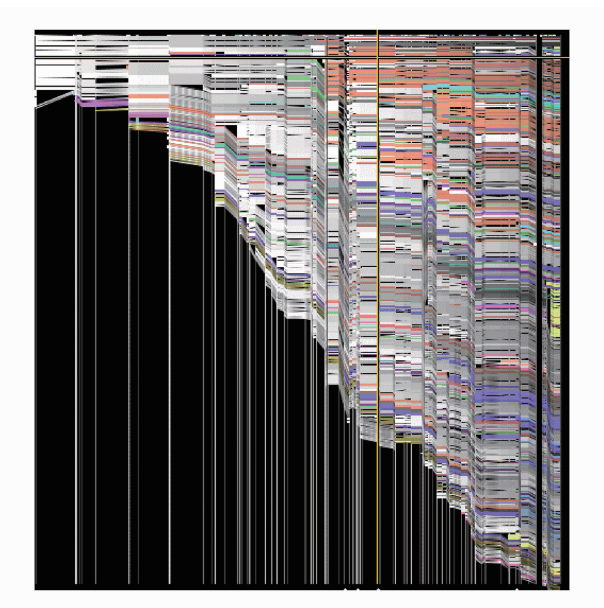

图 14 History Flow

还有一些研究通过动态变化呈现包含时间的文本信息[69-71]. TwitterScope[71] 以地图的形式呈现 Twitter 内容,并以颜色区分不同的主题,地图上的

① http://en.wikipedia.org

内容会随着时间动态地消失、融合. 图 15 所示 Streamit[69] 以动画的形式从左到右实时地呈现文本聚类的合并和分化.

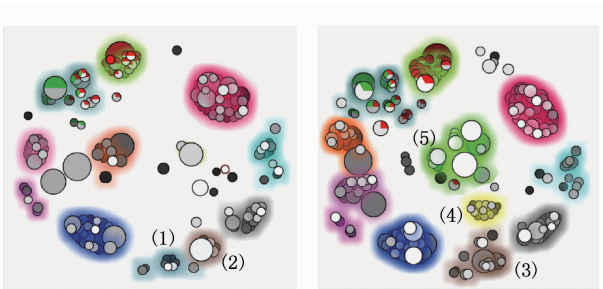

图 15　Streamit

此外，还有一些研究结合了时间和空间的信息. Thom 等[72] 的研究监测 Twitter 并自动将突发事件显示在地图上，图 16 所示 Whisper[73] 以向日葵为视觉隐喻，呈现信息在 Twitter 中的传播过程和规律；图 17 所示 TwitterMood[74] 以颜色表示心情，以呈现美国各州人们的情绪变化.

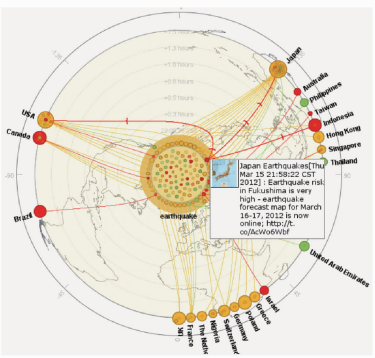

图 16　Whisper

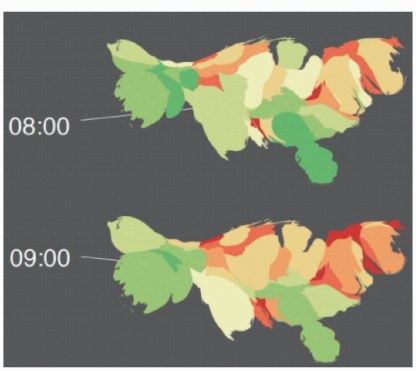

图 17　TwitterMood

包含时间的多层面信息可视化的主要难点在于如何有效地呈现多层的时间变量信息. 例如，如何在呈现单一属性随时间变化的同时，合理展示与时间相关的各个属性的关系.

3.3.2　其他基于多层面信息的可视化

Parallel Tag Cloud[75] 结合了标签云和常用于多维数据展示的平行坐标轴[76] 这 2 种可视化形式. 图 18 展示了 60 万个美国联邦上诉法院决策文档的关键词，每条平行坐标轴代表 13 个巡回上诉法院之一，直观地展示了各地区法案上的差异和联系.

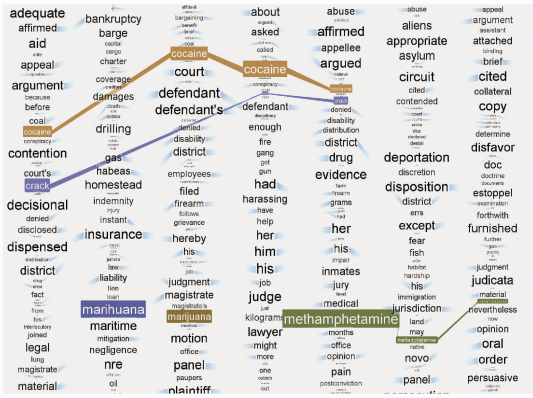

图 18　Parallel Tag Cloud

FacetAtlas[77] 同样用于呈现大量文本在多层面上的复杂关系，图 19 展示了对于 Google Health 文档中“糖尿病”的可视化，2 个大的聚类分别对应 I 型糖尿病和 II 型糖尿病，红色的连线代表相似的并发症，绿色的连线代表相似的症状.

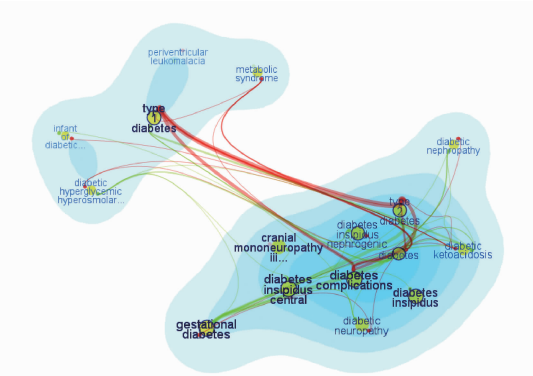

图 19　FacetAtlas

在需要通过视觉对比对文本信息进行分析和理解的情况下，与使用缩放等交互手段相比，能够同时提供多种呈现形式的可视化方案可以获得更好的认知效果[78]. Jigsaw[79]，FeatureLens[80] 和 ASE[19] 通过协同展示多个视角，呈现了文本集合中的多维度信息，便于用户对文本信息有全面深入的理解.

3.4　小　结

表 1 以主要的文本分析技术为线索，从可视化形式、可视化对象、是否包含时变数据、发表年几个方面对本文重点提到的文本可视化研究做了一个小结.

表 1 文本可视化方案小结

方案名称	主要文本分析技术	可视化形式	可视化对象	时变	发表年
NewsMap①	标注统计	TreeMap	新闻文本层次关系	否	2004
Map of the Market[40]	标注统计	TreeMap	股市数据层次关系	否	1999
FanLens[15]	标注统计	径向空间填充	命名实体层次关系	否	2008
LifeLines[48]	标注统计	时间轴	个人历史	是	1996
NameVoyager[55]	标注统计	叠式图	取名趋势	是	2005
BookVoyager[56]	标注统计	叠式图	图书销售趋势	是	2006
ThemeRiver[57]	标注统计	叠式图	主题的动态	是	2000
History Flow[66]	标注统计	叠式图	文本版本的变化	是	2004
Weber 等[68]	标注统计	2D 及 3D 螺旋图	周期性数据	是	2001
DocuBurst[16]	辞典匹配	径向空间填充	关键词及其上下位关系	否	2009
Wordle[7]	词袋模型	标签云	关键词	否	2009
Tagxedo②	词袋模型	特殊轮廓的标签云	关键词	否	2010
ManiWordle[8]	词袋模型	标签云	关键词	否	2010
Tilebars[10]	词袋模型	柱状图	词汇分布	否	1995
Contexter[11]	词袋模型	网络图	命名实体同现关系	否	2004
SparkClouds[50]	词袋模型	标签云及折线图	关键词的动态	是	2010
Lohmann 等[51]	词袋模型	带标记的标签云	关键词的动态	是	2012
Nguyen 等[52]	词袋模型	带标记的标签云	关键词的动态	是	2011
Tag River[62]	词袋模型	叠式图及标签云	关键词的动态	是	2011
Hassan-Montero 等[22]	词袋模型、模式匹配	标签云	关键词及其语义关系	否	2006
FeatureLens[80]	词袋模型、模式匹配	折线图及柱状图等	多层面信息	否	2007
Visual Backchannel[60]	词袋模型、词干提取	叠式图	主题的动态	是	2010
Document Card[9]	词袋模型、卡方分布[81]	标签云及图片	关键词、图	否	2009
TimeMines[47]	词袋模型、卡方分布	时间轴	时序新闻中的主体	是	2000
Parallel Tag Clouds[75]	词袋模型、对数似然[81]、辞典匹配	平行坐标轴及标签云	多层面信息	否	2009
Jigsaw[79]	词袋模型、命名实体识别	网络图、文本地图及散点图等	多层面信息	否	2008
FacetAtlas[77]	词袋模型、命名实体识别、Lucene③	文本密度图	多层面信息	否	2010
Phrase Net[14]	词袋模型、命名实体识别、模式匹配	网络图	命名实体语义关系	否	2009
Whisper[73]	词袋模型、情感分析	圆形地图	Twitter 信息的时空传播	是	2012
NewsRiver[59]	词性标注、命名实体识别	叠式图	主题的动态	是	2007
LifeFlow[49]	聚类	时间轴	事件序列	是	2011
Thom 等[72]	聚类(k-means)、Lucene	真实地图	突发事件时空监测	是	2012
InfoSky[42]	层次聚类	文本地图	文本集语义层次关系	否	2003
Paulovich 等[44]	层次聚类(HiPP)	文本地图	文本集语义层次关系	否	2008
ThemeCrowds[41]	层次聚类、词袋模型	TreeMap、标签云	Twitter 主题层次关系及其关键词	是	2011
MemeTracker[58]	特征统计、短语聚类	叠式图	主题的动态	是	2009
Galaxies[30]	文本特征映射	文本地图	文本集语义关系	否	1995
Galaxy[34]	文本特征映射	文本地图	网页集的语义关系	否	2002

① http://newsmap.jp/
② http://www.tagxedo.com/
③ http://lucene.apache.org/

续表 1

方案名称	主要文本分析技术	可视化形式	可视化对象	时变	发表年
Document Atlas[35]	文本特征映射(LSI、MDS)	文本密度图	文本集语义关系	否	2005
ProjCloud[36]	文本特征映射(LSP[82])、聚类(k-means)	文本地图及标签云	文本集语义关系及其关键词	否	2012
incBoard[70]	文本特征映射(MDS)、随机采样	网格图	文本集合的动态	是	2010
Text Map Explorer[29]	文本特征映射(ProjClus)、TFIDF	文本地图	文本集语义关系	否	2006
ET-Map[31]	文本特征映射(SOM)	文本地图	文本集语义关系	否	1996
Lin[33]	文本特征映射(SOM)	文本地图	文本集语义关系	否	1992
Skupin[43]	文本特征映射(SOM)	文本地图	文本集语义层次关系	否	2002
Kohonen 等[32]	文本特征映射(基于 SOM)	文本地图	文本集语义关系	否	2000
EventRiver[65]	文本特征映射、层次聚类(ROCK[83])	河流	事件的动态	是	2012
Streamit[69]	主题模型(LDA)、动态聚类	文本地图	主题集合的动态	是	2012
TIARA[61]	主题模型(LDA)、基于TFIDF 的算法	叠式图及标签云	主题内容的变化	是	2010
TwitterScope[71]	主题模型(LDA)、文本特征映射(MDS)	地图	Twitter 主题的动态	是	2012
TextFlow[64]	主题模型(基于 LDA)	河流	主题的细致演变	是	2011
ASE[19]	文本摘要	网络图及柱状图等	科技文献的多层面信息	否	2011
NETSPEAK[13]	概率检索	树	上下文关系	否	2011
Cui 等[53]	信息熵、上下文消歧[84]	带标记的标签云及折线图	关键词的动态	是	2010
CiteSpace[18]	突发检测算法[85]	网络图	文献共引关系	否	2006
WordTree[12]	后缀树	后缀树	上下文关系	否	2008
Zhang 等[20]	FP-tree	后缀树	文献共引关系	否	2009
TwitterMood[74]	情感分析	密度地图	情感的时空动态	是	2010

通过表 1 可以发现，文本地图常用于呈现聚类关系，叠式图常用于呈现时变信息，图 20 所示为对常见文本分析技术和可视化形式的关系的一个简要概括. 我们还可以发现，虽然情感分析是文本分析的研究热点，但对于文本情感信息的可视化研究相对较少，而对于文本时变信息的可视化近年来得到了较多研究和关注.

表 2 所示为对常见可视化形式的核心算法的一个小结.

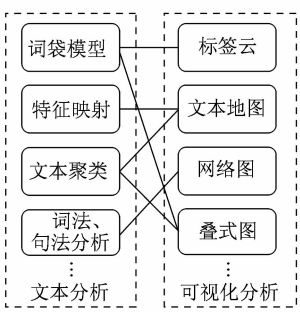

图 20　常见的文本分析技术及可视化形式

表 2　常见可视化形式的核心算法

可视化形式	核心算法
标签云(Wordle)	将所有文字降权排序，对于每一个文字首先将其随机放置在中心线附近，如果其与已经放置的文字重合，则以从内到外的螺旋线路径继续检测是否与已有文字重合，直到将其成功放置在空白区域. 在重合测试中，递归地将词语的边界框分成更小的矩形，使得较小的文字能够嵌在较大文字的空隙中.
文本地图	根据文本和聚类的相似度决定点与点、点集合与点集合的距离，相似度越高，它们在平面图上距离越近.
TreeMap	在绘制出最外层的矩形即树结构的根节点后，递归地细分内部空间成矩形块，即每一层节点的子节点都递归地划分各自的父节点. 各个矩形块的面积由各自的权重比例决定.

续表 2

可视化形式	核心算法
ThemeRiver	首先根据每个主题在离散时间上的权值进行插值,插值函数需满足在极值点导数为零的约束条件;然后进行叠式图的绘制:令主题 j 在 i 时刻的值为 f_{ji},$f_{\cdot i}=\frac{1}{2}\sum_{j=1}^{m}f_{ji}$,$\tilde{f}_{ji}=\sum_{j'=1}^{j}f_{j'i}-f_{\cdot i}$,$\tilde{f}_{0i}=-f_{\cdot i}$,$\tilde{f}_{ji}$ 和 $\tilde{f}_{j+1,i}$ 之间的宽度则代表了主题 i 的权重.

4 文本可视化技术的评价

与传统的文本分析技术不同,文本可视化技术的评价较少且难以利用精确的数据定量描述.目前,对于文本可视化技术,比较常见的评价方式有可用性测试(usability testing)、可用性检查(usability inspection)、个案研究(case study)以及对比试验.

4.1 可用性测试

可用性测试主要通过让用户使用可视化技术的原型或者系统,获取用户反馈的问题和意见,以总结和提出改善可视化技术的方案.可用性测试以用户为中心,能够在测试中直接了解实际用户怎样运用原型或系统获取信息,进而评价和改进原型或系统;因此它具有重要的设计指导意义,同时也是人机交互研究的重要研究内容.

4.2 可用性检查

可用性检查是指相关领域的专家学者运用自身较为丰富的经验,在应用场景中使用一系列方法检查可视化技术,发现可视化设计中存在的可用性问题.可用性检查最初在人机交互领域用于评价用户界面.与可用性测试不同,可用性检查不涉及到实际的普通用户,因此通常运用于可视化技术的原型测试阶段[86].

4.3 个案研究

个案研究是对一个应用场景进行集中、深入、全面分析的过程[87].不同于可用性检查,个案研究不必完全遵循严格的规则来检查每一个技术细节,而是深入地、长时间持续地关注于一个实例或事件.在个案研究中,研究者通过系统地查看事件、收集数据、分析信息,以期能够深刻地理解认知规律,从而明确未来研究的方向和重点.对于精心设计的、深入的个案研究,通常即使只有一个个案也能很好地揭示科学问题[88].

4.4 对比试验

对比试验通常是在相同的任务下研究者将待评价的可视化技术与类似应用场景下被广泛使用的可视化技术进行对比,其包含主观感受和客观数据.

5 结 语

文本可视化集成了文本分析、数据挖掘、数据可视化、计算机图形学、人机交互、认知科学等学科的理论和方法,结合了计算机的计算能力和人的认知能力,使得人们得以在与日俱增的海量文本中找到感兴趣的有用信息.

从本文的介绍可以看出,文本可视化技术已经取得丰富的成果,但文本可视化研究仍然处于起步阶段,面临诸多的挑战和问题:

1) 文本可视化技术需要具备良好的可扩展性,能够适应于不同的文本数据量大小、数据格式和数据质量.

2) 文本可视化仍然缺乏有效的研究范式来确保视觉呈现准确直接、交互形式符合直觉.近年来,虽然不乏新奇的文本可视化技术,但是极少像标签云一样被广泛接受和使用.

3) 文本可视化需要有效处理海量时变数据,目前的文本分析结果和效率仍然不尽人意,这将直接影响视觉呈现的可靠性和交互的平滑性.

4) 文本可视化仍然缺乏系统有效的评价标准,这主要是由于文本数据的多样性和复杂性造成的.

总之,文本可视化技术已经崭露头角,显示了其强大的信息呈现能力和作用,而目前面临的挑战主要来自于如何有效地整合各学科领域及相关知识,尤其是数据可视化和文本分析 2 个领域的技术,实现实用有效的可视化方案,以在文本数据日益增长的情况下,满足人们方便快捷地挖掘所需文本信息的要求.

参考文献(References):

[1] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620

[2] Haroz S, Whitney D. How capacity limits of attention influence information visualization effectiveness [J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2402-2410

[3] Heer J, Robertson G G. Animated transitions in statistical data graphics [J]. IEEE Transactions on Visualization and Computer Graphics, 2007, 13(6): 1240-1247

[4] Lamping J, Rao R, Pirolli P. A focus + context technique based on hyperbolic geometry for visualizing large hierarchies [C] //Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 1995: 401-408

[5] Sebastiani F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1): 1-47

[6] Viegas F B, Wattenberg M. TIMELINES: tag clouds and the case for vernacular visualization [J]. Interactions, 2008, 15(4): 49-52

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

[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] Strobelt H, Oelke D, Rohrdantz C, *et al*. Document cards: a top trumps visualization for documents [J]. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6): 1145-1152

[10] Hearst M A. TileBars: visualization of term distribution information in full text information access [C] //Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 1995: 59-66

[11] Grobelnik M, Mladenić D D. Visualization of news articles [J]. Informatica, 2004, 28(4): 375-380

[12] Wattenberg M, Viegas F B. The word tree, an interactive visual concordance [J]. IEEE Transactions on Visualization and Computer Graphics, 2008, 14(6): 1221-1228

[13] Riehmann P, Gruendl H, Froehlich B, *et al*. The netspeak wordgraph: visualizing keywords in context [C] //Proceedings of IEEE Pacific Visualization Symposium. Washington, D C: IEEE Computer Society Press, 2011: 123-130

[14] Van Ham F, Wattenberg M, Viegas F B. Mapping text with phrase nets [J]. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6): 1169-1176

[15] Lou X, Liu S, Wang T. FanLens: a visual Toolkit for dynamically exploring the distribution of hierarchical attributes [C] //Proceedings of IEEE Pacific Visualization Symposium. Washington, D C: IEEE Computer Society Press, 2008: 151-158

[16] Collins C, Carpendale S, Penn G. DocuBurst: visualizing document content using language structure [J]. Computer Graphics Forum, 2009, 28(3): 1039-1046

[17] Muller G A. Wordnet: a lexical database for English [J]. Communications of the ACM, 1995, 38(11): 39-41

[18] Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature [J]. Journal of the American Society for Information Science and Technology, 2006, 57(3): 359-377

[19] Dunne C, Shneiderman B, Gove R, *et al*. Rapid understanding of scientific paper collections: integrating statistics, text analysis, and visualization [R]. College Park, MD: University of Maryland, Human-Computer Interaction Lab, 2011

[20] Zhang J, Chen C, Li J. Visualizing the intellectual structure with paper-reference matrices [J]. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6): 1153-1160

[21] Han J, Pei J, Yin Y, *et al*. Mining frequent patterns without candidate generation: a frequent-pattern tree approach [J]. Data Mining and Knowledge Discovery, 2004, 8(1): 53-87

[22] Hassan-Montero Y, Herrero-Solana V. Improving tag-clouds as visual information retrieval interfaces [OL]. [2012-12-31]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.9998&rep=rep1&type=pdf

[23] 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

[24] Pearson K. On lines and planes of closest fit to systems of points in space [J]. Philosophical Magazine Series, 1901, 6(2): 559-572

[25] Greenacre M. Correspondence analysis in practice [M]. Boca Raton: Chapman & Hall/CRC, 2007

[26] Torgerson W S. Theory and methods of scaling [M]. New York: Wiley, 1958

[27] Kohonen T, Honkela T. Kohonen network [J]. Scholarpedia, 2007, 2(1): 1568

[28] Lin X. Map displays for information retrieval [J]. Journal of the American Society for Information Science, 1997, 48(1): 40-54

[29] Paulovich F V, Minghim R. Text map explorer: a tool to create and explore document maps [C] //Proceedings of the 10th International Conference on Information Visualization. Washington, D C: IEEE Computer Society Press, 2006: 245-251

[30] Wise J A, Thomas J J, Pennock K, *et al*. Visualizing the non-visual: spatial analysis and interaction with information from text documents [C] //Proceedings of International Conference on Information Visualization. Washington, D C: IEEE Computer Society Press, 1995: 51-58

[31] Chen H, Schuffels C, Orwig R E. Internet categorization and search: a self-organizing approach [J]. Journal of Visual Communication and Image Representation, Special Issue on Digital Libraries, 1996, 7(1): 88-102

[32] Kohonen T, Kaski S, Lagus K, *et al*. Self organization of a massive document collection [J]. IEEE Transactions on Neural Networks, 2000, 11(3): 574-585

[33] Lin X. Visualization for the document space [C] // Proceedings of IEEE Conference on Visualization. Washington, D C: IEEE Computer Society Press, 1992: 274-281

[34] Whiting M, Cramer N. WebTheme™: understanding web information through visual analytics [M] //Lecture Notes in Computer Science, Heidelberg: Springer, 2002, 2342: 460-468

[35] Fortuna B, Mladenić D, Grobelnik M. Visualization of text document corpus [J]. Informatica, 2005, 29(4): 497-504

[36] 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

[37] Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques [OL]. [2012-12-31]. http: // www. cs. cmu. edu/~dunja/KDDpapers/Steinbach_IR. pdf

[38] Skupin A. From metaphor to method: cartographic perspectives on information visualization [C] //Proceedings of IEEE Conference on Symposium on Information Visualization. Washington, D C: IEEE Computer Society Press, 2000: 91-97

[39] Johnson B, Shneiderman B. Tree-maps: a space-filling approach to the visualization of hierarchical information structures [C] //Proceedings of IEEE Conference on Visualization. Los Alamitos: IEEE Computer Society Press, 1991: 284-291

[40] Wattenberg M. Visualizing the stock market [C] // Proceedings of ACM SIGCHI Extended Abstracts on Human Factors in Computing Systems. New York: ACM Press, 1999: 188-189

[41] Archambault D, Greene D, Cunningham P A D, *et al*. ThemeCrowds: multiresolution summaries of twitter usage [C] //Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents. New York: ACM Press, 2011: 77-84

[42] Kienreich W, Sabol V, Granitzer M, *et al*. Infosky: a system for visual exploration of very large, hierarchically structured knowledge spaces [OL]. [2012-12-31]. http: //www. kde. cs. uni-kassel. de/ws/LLWA03/fgwm/Resources/FGWM03_02_ Wolfgang_Kienreich. pdf

[43] Skupin A. A cartographic approach to visualizing conference abstracts [J]. IEEE Computer Graphics and Applications, 2002, 22(1): 50-58

[44] Paulovich F V, Minghim R. HiPP: a novel hierarchical point placement strategy and its application to the exploration of document collections [J]. IEEE Transactions on Visualization and Computer Graphics, 2008, 14(6): 1229-1236

[45] Fruchterman T M J, Reingold E M. Graph drawing by force-directed placement [J]. Software: Practice and Experience, 1991, 21(11): 1129-1164

[46] Klein J L. Statistical visions in time: a history of time series analysis [M]. Cambridge UK: Cambridge University Press, 1997: 1662-1938

[47] Swan J, Jensen D. TimeMines: Constructing timelines with statistical models of word usage [OL]. [2012-12-31]. http: // www. cs. cmu. edu/~dunja/KDDpapers/Swan_TM. pdf

[48] Plaisant C, Milash B, Rose A, *et al*. LifeLines: visualizing personal histories [OL]. [2012-12-31]. http: //hcil2. cs. umd. edu/trs/95-15/95-15. html

[49] Wongsuphasawat K, Guerra G O, Mez J A, *et al*. LifeFlow: visualizing an overview of event sequences [C] //Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2011: 1747-1756

[50] 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

[51] Lohmann S, Burch M, Schmauder H O R, *et al*. Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds [C] //Proceedings of International Working Conference on Advanced Visual Interfaces. New York: ACM Press, 2012: 753-756

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

[53] Cui W W, Wu Y C, Liu S X, *et al*. Context preserving dynamic word cloud visualization [C] //Proceedings of IEEE Pacific Visualization Symposium. Los Alamitos: IEEE Computer Society Press, 2010: 121-128

[54] Byron L, Wattenberg M. Stacked graphs-geometry & aesthetics [J]. IEEE Transactions on Visualization and Computer Graphics, 2008, 14(6): 1245-1252

[55] Wattenberg M. Baby names, visualization, and social data analysis [C] //Proceedings of IEEE Symposium on Information Visualization. Washington, D C: IEEE Computer Society Press, 2005: 1-7

[56] Wattenberg M, Kriss J. Designing for social data analysis [J]. IEEE Transactions on Visualization and Computer Graphics, 2006, 12(4): 549-557

[57] Havre S, Hetzler B, Nowell L. ThemeRiver: visualizing theme changes over time [C] //Proceedings of IEEE Symposium on Information Visualization. Washington, D C: IEEE Computer Society Press, 2000: 115-123

[58] Leskovec J, Backstrom L, Kleinberg J. Meme-tracking and the dynamics of the news cycle [C] //Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009: 497-506

[59] Ghoniem M, Luo D, Yang J, *et al*. NewsLab: exploratory broadcast news video analysis [C] //Proceedings of IEEE Symposium on Visual Analytics Science and Technology. Washington, D C: IEEE Computer Society Press, 2007: 123-130

[60] Dork M, Gruen D, Williamson C, *et al*. A visual backchannel for large-scale events [J]. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 1129-1138

[61] Wei F, Liu S, Song Y, *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: ACM Press, 2010: 153-162

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

[63] Forbes A G, Alper B, H O Llerer T, *et al*. Interactive folksonomic analytics with the tag river visualization [OL]. [2012-12-31]. http: //vialab. science. uoit. ca/textvis2011/papers/textvis%202011-forbes. pdf

[64] Cui W, Liu S, 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

[65] Luo D, Yang J, Krstajic M, *et al*. EventRiver: visually exploring text collections with temporal references [J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(1): 93-105

[66] Viegas F B, Wattenberg M, Dave K. Studying cooperation and conflict between authors with history flow visualizations [C] //Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2004: 575-582

[67] Carlis J V, Konstan J A. Interactive visualization of serial periodic data [C] //Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology. New York: ACM Press, 1998: 29-38

[68] Weber M, Alexa M, Muller W. Visualizing time-series on spirals [C] //Proceedings of IEEE Symposium on Information Visualization. Los Alamitos: IEEE Computer Society Press, 2001: 7-13

[69] Alsakran J, Chen Y, Luo D, *et al*. Real-time visualization of streaming text with a force-based dynamic system [J]. IEEE Computer Graphics and Applications, 2012, 32(1): 34-45

[70] Pinho R D, Oliveira M D, Andrade Lopes A D. An incremental space to visualize dynamic data sets [J]. Multimedia Tools and Applications, 2010, 50(3): 533-562

[71] Gansner E R, Hu Y, North S C. Visualizing streaming text data with dynamic maps [OL]. [2012-12-31]. http://arxiv.org/abs/1206.3980

[72] Thom D, Bosch H, Koch S, *et al*. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages [C] //Proceedings of IEEE Conference on Pacific Visualization Symposium. Los Alamitos: IEEE Computer Society Press, 2012: 41-48

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

[74] Mislove A, Lehmann S, Ahn Y Y, *et al*. Pulse of the nation: US mood throughout the day inferred from twitter [OL]. [2012-12-31]. http: //www. infosthetics. com/archives/2010/07/pulse_of_the_nation_us_mood_throughout_the_day_inferred_from_twitter. html

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

[76] Inselberg A, Dimsdale B. Parallel coordinates: a tool for visualizing multi-dimensional geometry [C] //Proceedings of IEEE Visualization Conference. Los Alamitos: IEEE Computer Society Press, 1990: 361-378

[77] Cao N, Sun J, Lin Y, *et al*. FacetAtlas: multifaceted visualization for rich text corpora [J]. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 1172-1181

[78] Plumlee M D, Ware C. Zooming versus multiple window interfaces: cognitive costs of visual comparisons [J]. ACM Transactions on Computer-Human Interaction, 2006, 13(2): 179-209

[79] Stasko J, Görg C, Liu Z. Jigsaw: supporting investigative analysis through interactive visualization [J]. Information Visualization, 2008, 7(2): 118-132

[80] Don A, Zheleva E, Gregory M, *et al*. Discovering interesting usage patterns in text collections: integrating text mining with visualization [C] //Proceedings of ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2007: 213-222

[81] Cressie N, Read T R C. Multinomial goodness-of-fit tests [J]. Journal of the Royal Statistical Society. Series B (Methodological), 1984, 46(3): 440-464

[82] Paulovich F V, Nonato L G, Minghim R, *et al*. Least square projection: a fast high-precision multidimensional projection technique and its application to document mapping [J]. IEEE Transactions on Visualization and Computer Graphics, 2008, 14(3): 564-575

[83] Guha S, Rastogi R, Shim K. ROCK: a robust clustering algorithm for categorical attributes [J]. Information systems, 2000, 25(5): 345-366

[84] Schütze H. Automatic word sense discrimination [J]. Computational Linguistics, 1998, 24(1): 97-123

[85] Kleinberg J. Bursty and hierarchical structure in streams [J]. Data Mining and Knowledge Discovery, 2003, 7(4): 373-397

[86] Nielsen J. Usability inspection methods [C] //Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 1994: 413-414

[87] Flyvbjerg B. Case study [M]. Denzin N K, Lincoln Y S, 4th ed. The Sage Handbook of Qualitative Research. Thousand Oaks: Sage Publication, 2011: 301-316

[88] Flyvbjerg B. Five misunderstandings about case-study research [J]. Qualitative Inquiry, 2006, 12(2): 219-245