



基于中文文本的可视化技术研究

袁海¹, 陈康², 陶彩霞², 陈翀²

(1. 中国电信股份有限公司江苏分公司 南京 210037;

2. 中国电信股份有限公司广东研究院 广州 510630)

摘要: 文本可视化是数据可视化的新兴主题,其目的是让人直观地认识数据,敏捷地发现数据中隐藏的规律。首先介绍了文本可视化的过程,包括信息收集、预处理、视觉呈现和交互功能等;然后通过分析和使用各种中文文本可视化的工具,从基于标签云、树图、关联和时间序列等多方面研究可视化技术;最后对比了这几种技术的差异和优缺点,并指出各技术的适用场景。文本可视化能够更好地帮助用户处理日益增长的文本信息,提高数据的可视化效率和转换成利益的效率。

关键词: 文本分析;信息可视化;文本可视化;交互界面

doi: 10.3969/j.issn.1000-0801.2014.04.017

Research on Visualization Techniques Based on Chinese Texts

Yuan Hai¹, Chen Kang², Tao Caixia², Chen Chong²

(1. Jiangsu Branch of China Telecom Co., Ltd., Nanjing 210037, China;

2. Guangdong Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China)

Abstract: The text visualization process was introduced at first, including information collection, preprocessing, visual presentation and interactive functions. Then based on the analysis of various Chinese text visualization tools, the visualization technology areas such as multi-based tag cloud, tree map, association and time sequences were researched. At last, the advantages and disadvantages of these techniques were compared from a few aspects and the various technologies applicable to the scene were pointed out. Text visualization can help users to better deal with the growing text information to improve the efficiency of data visualization and converted into efficiency benefits.

Key words: text analysis, information visualization, text visualization, interaction interface

1 引言

最近几年,随着博客、社交网络等互联网产品的兴起,海量信息不断涌现,从这些信息中提取、分析、发现重要信息的方法也日新月异。这些技术中,文本可视化技术是发展比较快的一种。诸如传统的书籍、论文,最近的 Web 文本、微博博文等,其内容都是以文本为载体的,而文本可视化则是根据这些文本的内容,挖掘、提炼出有用的信息,并且以带有交互功能的图形、图表等形式展示给用户,并不

是将大批量、大段的文字简单地返回。文本可视化已经渗透到计算机以外的各个学科,如科学、艺术、社会学等,属于一个多样化、边缘化的技术。在讨论文本可视化分析的内容时,前人按照不同的研究给出了不同的分类结果,参考文献[1]给出了文本可视化研究的如下 3 个层次:

- 基于词频统计的文本可视化;
- 基于聚类算法的文本可视化;
- 基于语义的文本可视化。

文本可视化所要做的工作不仅是将文字转换成几个

图形、图表,更大的作用在于发现一篇文档内容的结构、主题和隐含的特征,或者发现一系列文档中存在的关系、模型等。同时以简洁、有交互性的图形展示给使用者,即将枯燥的信息与数据转换成美丽的、能给人深刻印象并且有意义的图形。只有将数据和信息用图形和图像表示出来,才能展现隐藏的知识。而文本可视化能够加快数据处理的速度,使结构化、非结构化的数据得以充分利用;可以在人、数据之间建立通信,从而直观观察数据中隐含的规律、热点,为发现定律提供有力的工具;可以实现对计算和编程过程的引导和控制,通过交互手段改变呈现结构和效果,并观察和比较其影响。

2 文本可视化的方法与相关工具

文本可视化分析技术的研究方向有基于文本内容、基于文本关系、基于多层面信息以及对于文本流的可视化分析。根据文本可视化的设计方案和不同的应用目的,文本可视化可以分为以下几类:基于词汇的可视化、基于篇章的可视化、基于时间序列的可视化和基于主题领域的可视化。本文的可视化研究主要是对中文文本内容进行研究。

一般来说,文本可视化可以包括以下几个过程:信息收集、预处理、知识表示、视觉呈现和交互等。其中,信息收集是文本可视化的资料收集阶段,是许多数据可视化必备的;而预处理则是对文本进行的各种操作,包括分词、去停用词、提取关键词等自然语言处理;知识表示则是根据要求将文字组织成某种形式的过程;视觉呈现则是选择适合的图形或图表将可视化文本呈现出来的过程,需要考虑的因素包括文本的数量、元素和效果等;交互则是指人机交互,添加用户对数据图形的操作,实现互动、动态展示,是最近研究文本可视化的热点。

文本预处理过程是一个自然语言处理的过程,其文本分析关键技术包括分词、命名实体识别、关键词抽取、主题分析以及情感分析等。其中,分词是很重要的一环,分词的准确率会对后面的分析的正确性产生很大的影响。相对于英文的分词,中文的分词难度大很多,因为中文的字并没有像英文那样用空格分开每个单词,这也是造成中文分词困难的最大原因。然而随着技术的成熟,中文分词取得了很大的成功。目前,在中文分词方面,做得比较好的分词器有NLPIR汉语分词系统^[2]、哈尔滨工业大学语言技术平台^[3]、斯坦福大学的自然语言处理套件^[4]、复旦大学的分词器(FudanNLP)^[5],这几个分词器都有很高的准确率和速度,

而且功能齐全,被许多机构、学校以及企业使用。

在分词后,最直接的文本可视化就是基于词汇的可视化,可以分为3种:基于命名实体的可视化、基于词频的可视化和基于词汇分布的可视化。命名实体是指一些关键的人名、地名、术语等词语,这些词语一般具有特殊的含义和特定的关系,因此将这些词语进行可视化能够了解一篇文档的主语关系;文本可视化工具Contexter^[6]就是一个将命名实体可视化的自动化软件;基于词频的可视化是在统计文本词语的出现频次后,根据词频赋予每个词语不同的权重,并在前端将这些词语重新排列展示。其中,最常用的词频计算方法是TFIDF^[7],而最常用的可视化形式是标签云(或称为词云)^[8],文本可视化软件WordLe是其中广泛采用的代表之一^[9]。基于词汇分布的可视化,是反映词汇在文本中的分布情况,通常是将整文按段落、章节等划分文本,能够使用户查询词语以及词间的关系。在这方面Tilebars^[10]就是反映词汇分布的可视化软件。此外,还有许多技术或软件用于对词汇的可视化,如结合主题特征的TopicIslands^[11]、添加时间维度的ThemeRiver^[12]、动态变化的MemeTracker^[13]以及主题分类的NewsMap^[14]等。

对单篇文本进行可视化的另一种方法是基于文本关系的可视化,也可以称为基于篇章内容的文本可视化。基于篇章,即更多的是关注、发现文章中特定的语义关系,展示文章的主题和核心内容,呈现文章脉络和阅读线索。这类可视化一般是将文本通过特征分析后再展示,特征分析的方法有很多,如主成分分析、多维尺度分析、潜在语义分析以及主题模型。在这方面,常见的可视化软件或方法有Phrase Net^[15]、DocuBurst^[16]、WordTree^[17]等。此外,研究者对基于文本外在关系的可视化研究也有诸多进展,主要包括两方面内容:多文本的词语关系可视化和文本间引文关系的可视化。多文本的词语关系可视化多用于信息检索、主题检测和话题演变等方面,这些技术可以是传统的,如分类、聚类,也可以是最近的新技术,如主题模型,其目的都在于挖掘大量文本信息的内在关系、关键主题;文本间引文关系的可视化,主要是研究文章作者、文章引用或会议等内容,以方便理解文献的发表情况和研究进展。在这方面,主要的展示工具有FP-Tree^[18]、ContextTour^[19]和FacetAtlas^[20]等。

文本的可视化呈现和交互是直接为用户相关的环节。在可视化呈现时,主要考虑的因素包括图形的选择、形状的选择、颜色的选择等。其中,图形的选择包括树状图、坐标轴、放射状图、网络图等;形状的选择则包括点的设计、



边的设计,具体有方形、圆形、星形等;颜色的选择则有千万种,一般来说只要能够区分不同词语和不同特征就能够满足,现实中偏向于选择鲜艳的颜色,如红色、蓝色等。另外,可视化字体的选择、大小、方位等也很重要。交互功能则是让用户参与到可视化展示中,通过需求呈现不同的文本特性。在文本可视化中,主要用到的交互方式包括高亮、缩放、更新、转换、标注、分层查看以及焦点加上下文等。

3 文本可视化技术应用研究

文本可视化是数据可视化的一个分支,国内外有很多成果,特别是技术发展以及计算机的普及,使得许多互联网企业、运营商都想将自己的数据可视化,立体地展示给用户,以获取更多的用户支持、广告投资。但随着社交方式的变迁,利用旧的可视化方法展示新的数据遇到了一些障碍,如数据流、动态数据等。同时,由于语言不同,可视化工具的底层设计、展示方式也不同,即不同语言的可视化工具不能够通用,需要研究符合本语言的可视化工具。本文根据前人的研究,对中文文本进行可视化研究,包括工具调研、设计、效果以及评价等。

3.1 基于标签云的文本可视化

标签云也称为词云、文字云,是词频可视化中最为典型的形式。标签云是关键词的视觉化描述,用于汇总用户生成的标签或一个网站的文字内容。标签一般是独立的词汇,常常按字母顺序排列,其重要程度又能通过改变字体大小或颜色表现,所以标签云可以灵活地依照字母顺序或热门程度检索一个标签。大多数标签本身就是超级链接,

直接指向与标签相联的一系列条目。标签云很适合作用来挖掘和展现热点,微博中的话题、用户标签都很适合用标签云展示。

标签云按作用的不同可以分为三大类:第1类用于描述网站的每个独立条目,第2类用于描述网站所有条目的标签情况,第3类是将标签作为一个数据项目的工具,用于表示整个集合中各项目的量。按形式分,可以分为平面型和球面型:平面型即2D型,一般是静态的,标签无链接;而球面型即3D型,标签是动态展示的,可以根据鼠标的移动而转动,并能够点击打开链接或搜索。研究表明,网站采用标签云的好处有:方便信息管理、促进协作分享、增加参与度等。

标签云的实现方案有很多,大部分是Web版的,如WordLe^[21]、ImageChef^[22]、Tagul^[23]等。这些工具不只是简单地将文字展示出来,更多的是让用户参与其中,实现个性化展示。例如,对微博的话题热点进行可视化,截取其中最频繁的150个词用WordLe展示,如图1所示。通过标签云的可视化,看出“官方推荐”这个词在这些话题中出现最多,而且人们都在讨论“HOT男人”、“生日”、“YOKA”等话题,而热点讨论的人物包括“华晨宇”、“刘忻”等。

除了Web端的标签云制作外,一些传统的工具也可以用来开发标签云,如R语言。R语言提供“wordcloud”分组用于制作标签云,其做法简单快捷,还可以与软件集成。如果需要对比表1的微博标签按性别展示成标签云,展示的结果如图2所示。对比表1和图2,发现通过标签云展示数据会更直观且更具可读性。



图1 微博话题的标签云可视化

表 1 微博用户标签(部分)

性别	标签
男	投资理财,教育就业,体育
女	健康,时尚,文艺,娱乐
女	“菇凉们”,轻熟女,网购控,衣服,包包,鞋子,护肤,逛街
男	新闻资讯,旅游
男	体育
女	八卦,美食
女	娱乐,旅游,美食,星座命理
女	高三,追星,可爱,爱睡觉,娱乐
男	美剧,睡觉,汽车,新闻,吃,好性格,唱歌,八卦,囤,“偶喜欢玩”
男	上网,时尚,“8090”,管理,音乐爱好者,饰品,服装,宁乡玉潭镇,民生无小事
男	书,诚信,爱,正能量,创业是我的梦想,创新,旅游
男	O 型血,陈奕迅,看小说,建筑,健康,美食,文艺,IT 数码,体育
女	小狮子,“90 后”,处女座
女	ModernFamily,runningman,美剧重症患者
女	听歌看电影,手机控,微博控,“睡觉 zzZZ”,表情帝,大脸妹,水瓶座,“90 后”,吃货
女	新闻资讯,美食
女	视频音乐,旅游,北京生活,微博奇葩,搞笑幽默

3.2 基于树图的文本可视化

树图也称为树状图、树形地图,是人机界面大师 Shneiderman B 发明的资料表示法^[24],主要目的是在一张图内一目了然整体状况,由各个元件量的大小决定图示大小,并具有群组功能。简单而言,树图是一种层次数据可视

化的方法,用一定面积的块表示数据中的个体,使用空间位置表示个体之间的关系,而块的大小则表示个体的权重。树图拥有与标签云相似的好处,都能够让用户快捷地对文本中的主题、关键词有所了解,区分关键词作用的大小,并能够将这些词语按一定的要求组织起来,有的还能提供交互功能,用户点击某层,就能够将该层的内容展开,以查看更多、更详细的数据。

树图的主要研究内容一般包括 3 个方面:基本布局算法、视觉设计、交互设计。基本布局算法有递归算法、贪心策略算法、由内向外布局算法等,这些算法适用于不同的场景,可以展示出不同的效果,如今研究更多的是多层次、多节点以及有周期性的布局算法。在选择这些布局算法时,需要考虑到算法的稳定性、可读性、连续性等性质。树图的视觉设计则是在选定布局算法后,在一些细节上进行修改,如标签的颜色显示,文字的大小、字体,还有块的纹理、投影效果等。树图的交互设计也是图形设计的重要一环,是提高用户体验很好的途径,树图的交互方法有改变权重、改变颜色、切换布局以及焦点上下文等。最近有许多新颖的设计,如 Maigic Lens^[25]、Semantic Zooming^[26]等。

树图的图形选择有很多,一般是矩形,还有圆形、心形、多角形、球形等,树图与柱状图结合还可以组成展示功能更好的可视化模型。用树图进行文本可视化有很多用途,如展示关键词、分析包含与逻辑关系等。用 D3^[27]树图展示的效果如图 3 所示,可以看出占用比例大的关键词。利用经典的 Treemap 布局^[28]展示文本的示例如图 4 所示,该



图 2 根据性别形成的标签云



工具提供多个布局面板、多种大小、颜色选择以及交互模块,还可以双击某块进入查看详细内容,是树图展示不错的选择。

3.3 基于关联的文本可视化

标签云和树状图都拥有直观和美观两大优点,但其展现能力也有局限性,如仅对关键词进行离散的展示,丢失了关键词之间的关联性和文本内容的紧凑性,也无法展现文章的主题,因此研究者探索利用其他方法来展示文本的多维度信息,基于 FacetAtlas 算法的可视化就是其中一种。参考文献[20]首先将 FacetAtlas 用于可视化,克服了简单投影造成的信息丢失,并将投影方法与点线图相结合,能够在同一个版面表达数据的多维信息,帮助理解文本潜在的

关系。这种方法还可以与文本聚类技术相结合,从而分析多文本的主题关系,并配以丰富的交互方式,传达更多的信息。

FacetAtlas 算法是一个力学布局算法,其利用力学原理,令节点互相排斥或互相吸引,基本思想是将图看成一个以钢环为顶点、以弹簧为边的物理系统,系统被赋予某个初始状态后,弹簧弹力(引力和斥力)的作用会导致钢环运动,直到系统总能量减少到最小值时停止。采用不同的力学原理会生成不同的布局结构,例如,可以用带电荷粒子的电力模拟节点的排斥力或用弹簧弹力模拟边的吸引力。使用 ForceAtlas 算法的目的是将有联系的节点紧靠在一起,并与其他节点明显分隔开,将数据形成一个个群组。

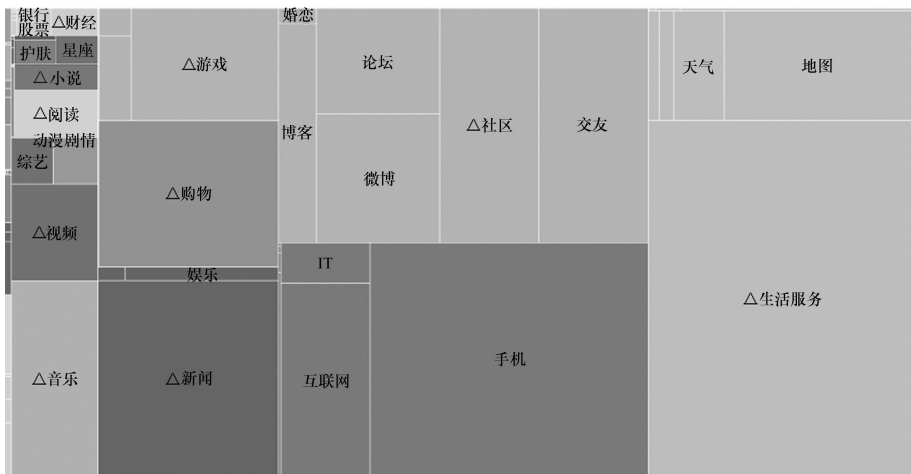


图3 D3 树图展示的效果

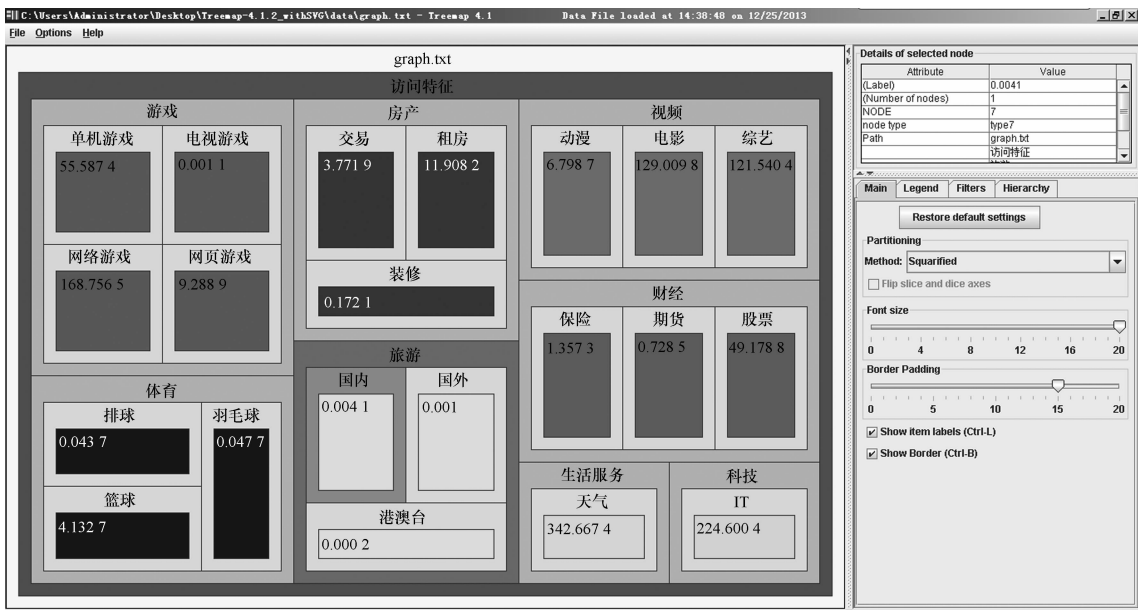


图4 文本的树图可视化

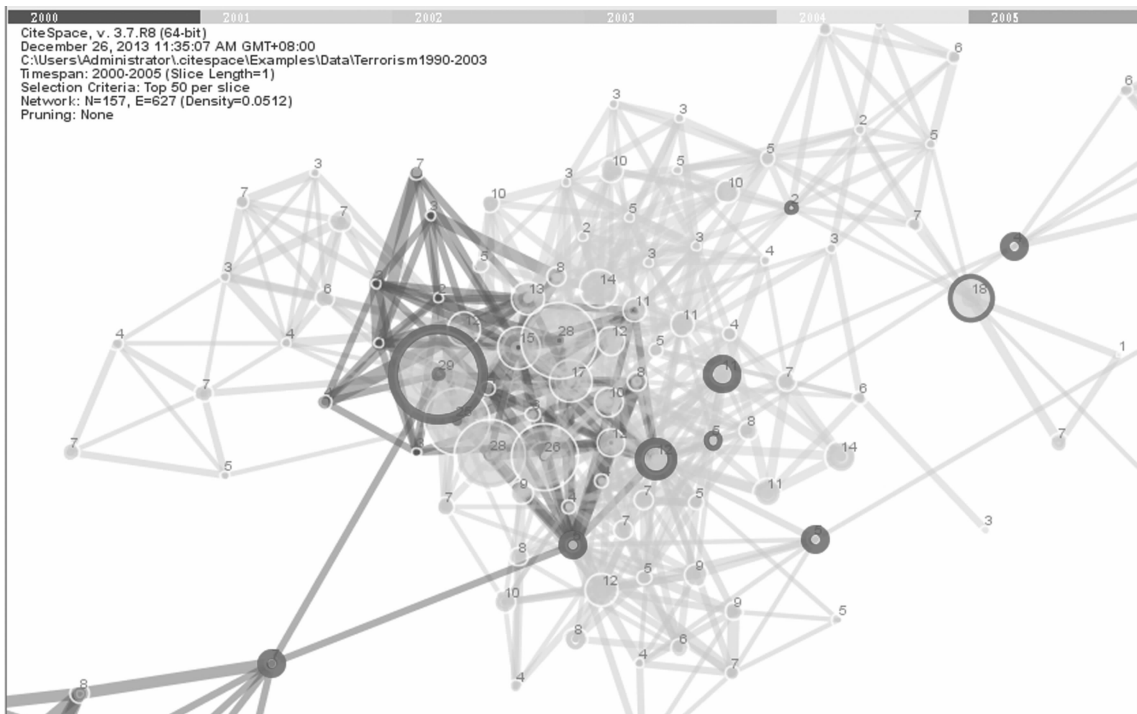


图6 CiteSpace 文献共现可视化效果

示文本信息,是指以动画或具有更新性能的展示方法,研究信息的传播、流动规律,如动态呈现 Twitter 信息的 TwitterScope^[34]、TwitterMood^[35]。将时间信息融入文本可视化的研究方向包括:添加时间轴、添加顺序时间河、添加螺旋形时间流、与空间信息结合等。针对文本本身,展示方式又可以分为按主题变化、按内容变化、按人物变化等。

基于时间的文本可视化也很关注展示的视觉效果,如使用不同的图形展示不同的元素、使用不同的颜色标注不同的主题、采用不同的块代表不同的功能。这些展示不仅让画面更吸引人,而且增加可读性,可以清晰地理解文本内容的变化。如今,很多可视化工具都集成了多种展示功能,如 TimeFlow^[36]。TimeFlow 能够以多种形式展现文本内容,如按时间流(timeline)方式、按日期(calendar)方式、按词频(bar graph)方式等,而且还提供多种交互功能,能够让用户选择展示的范围。TimeFlow 按时间展示关键词时,观察界面的关键词即可以理解文本内容的变化,将鼠标放在关键词上即可展示这个词语所在的文本名称。如图 7 所示是按时间查看文本的分布情况,时间能够精确到时分秒,这种展示很适合分析日志文件,如搜索引擎的点击日志。

3.5 文本可视化工具的比较

用于文本可视化的工具有很多,每一种工具都有自己的优点,能够在某一方面显式地展现文本的内容、结构、关

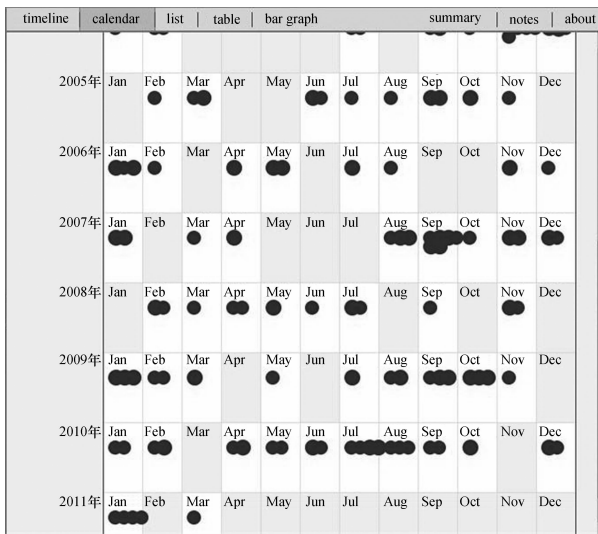


图7 按时间查看文本的分布

系,但它们使用的分析技术、设计理念和交互功能又有差异。从多个角度分析比较几种工具的应用场景,对比结果见表 2。从表 2 看,标签云的可视化形式适用于展示文本中的一系列关键词;树图由于其形状特征,能够用来展示文本主题的层次关系或主题关键词;关联关系则用到点、线,并用网状的形式表现文本内容的关系;结合时间的可视化则是将文本事件的序列进展以动态的形式呈现。这些工具各有各的好处,选择适应的场合能收到事半功倍的效果。

表2 文本可视化工具的比较

可视化技术	文本分析技术	可视化对象	交互功能	典型工具
标签云	词袋模型	关键词	无	Wordle、SparkClouds
树图	标注统计	数据层次关系	有	Treemap、NewsMap
关联关系	命名实体识别	多层面信息	有	FacetAtlas、CiteSpace
时间序列	流式统计	事件、时间序列	有	TimeMines、TimeFlow

4 结束语

文本可视化是一个多学科交叉的研究方向,其集成了文本分析、数据挖掘、人机交互等学科的理论与方法,还增加了人类心理学的概念,解决了文本可视化在图形绘制、图形展现和数据分析方面的众多问题;同时能更好地帮助用户处理日益增长的文本信息。未来可视化分析技术可改善现有布局算法,以提高实时的、大量的数据可视化效率,并且综合运用各种人际交互方法,提高展现效果,转化成可规模化商业应用的能力。本文通过调研各种中文文本可视化工具,分析各技术的分析方法、交互性能、使用对象等。

从这些工作来看,文本可视化研究取得了很丰富的成果,文本可视化技术已经被广泛使用。同时,随着情感分析、舆情分析技术的成熟,可视化技术将会进一步发展。但也有很多问题需在处理过程中加以考虑,如无缝融合信息可视化和数据分析技术,如何更好地处理海量、实时分布的互联网数据信息,这也将是文本可视化下一步的研究工作。有关社交网络的文本挖掘研究正蓬勃发展,未来将对现实世界产生巨大的作用。

参考文献

- Wise J A, Pennock K, Lantrip D, *et al.* Visualizing the non-visual: spatial analysis and interaction with information from text documents. Proceedings of information visualization, Atlanta, GA, USA, 1995:51~58
- <http://ictclas.nlpir.org/>,2014
- <http://www.ltp-cloud.com/>,2014
- <http://nlp.stanford.edu/software/index.shtml>,2014
- <http://code.google.com/p/fudannlp/>,2014
- Mladenec M G D. Visualization of news articles. <http://ailab.ijs.si/dunja/SiKDD2004/Papers/GrobelnikMladenec-Contexter.pdf>,2008
- Sebastiani F. Machine learning in automated text categorization. ACM Computer Surveys,2002,34(1):1~47
- Viegas F B, Wattenberg M. TIMELINES: tag clouds and the

- case for vernacular visualization. Interactions, 2008, 15(4):49~52
- Viegas F B, Wattenberg M, Feinberg J. Participatory visualization with word. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6):1137~1144
- Hearst M A. TileBars: visualization of term distribution information in full text information access. Proceedings of the SIGCHI Conference on Human Factors in Computing System, Austin, United States, 1995:59~66
- Miller M, Wong P, Brewster M, *et al.* TOPICISLANDS-a wavelet-based text visualization system. Proceedings of IEEE Visualization, North Carolina, USA, 1998: 189~196
- Havre S, Hetzler B, Nowell L. ThemeRiver: visualization theme changes over time. Proceedings of IEEE Symposium on Information Visualization, Salt Lake City, UT, USA, 2000:115~123
- Leskovec J, Backstrom L, Kleinberg J M. Meme tracking and the dynamics of the news cycle. Proceedings of KDD 2009, Paris, France, 2009:497~506
- <http://www.newsmap.jp>,2014
- Van Ham F, Wattenberg M, Viegas F B. Mapping text with phrase nets. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6):1169~1176
- Collins C, Carpendale S, Penn G. DocuBurst: visualizing document content using language structure. Computer Graphics Forum,2009,28(3):1039~1046
- Wattenberg M, Fernanda B. The word tree, an interactive visual concordance. IEEE Transactions on Visualization and Computer Graphics, 2008,14(6):1221~1228
- Han J, Pei J, Yin Y, *et al.* Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Mining and Knowledge Discovery, 2004, 8(1):53~87
- Lin Y R, Sun J, Cao N, *et al.* Contextour: contextual contour visual analysis on dynamic multi-relational clustering. Proceedings of SDM 2010, Columbus, Ohio, 2010
- Cao N, Sun J, Lin Y R, *et al.* Facetatlas: multifaceted visualization for rich text corpora. IEEE Transactions on Visualization Computer Graphic, 2010,16(6):1172~1181
- <http://www.wordle.net/>,2014
- http://www.imagechef.com/ic/word_mosaic/,2014
- <http://tagul.com/>,2014
- <http://www.cs.umd.edu/hcil/treemap-history/>,2014
- Jin L, Banks D C. TennisViewer:a browser for competition trees. IEEE Computer Graphics and Applications,1997,17(4):63~65
- Shi K, Irani P, Li B. An evaluation of content browsing technique: for hierachica space:filling visualization. Proceedings of IEEE Symposium on Information Visualization, Los Alamitos: IEEE Computer Society Press, 2005:81~88
- <http://d3js.org/>,2014
- Bederson B B, Shneiderman B, Wattenberg M. Ordered and