大数据可视化技术

学习目标

（1）掌握大数据可视化的基本概念、发展历程及其在大数据分析中的重要性。

（2）了解大数据可视化的主要分类，包括信息图形、信息可视化、科学可视化和统计图形，并理解它们之间的区别与联系。

（3）对文本可视化、网络可视化、时空数据可视化、多维数据可视化等特定类型的数据可视化技术进行深入研究，理解其原理、方法和应用场景。

（4）熟悉并掌握当前主流的大数据可视化工具，如 Tableau、ECharts、R 语言（ggplot2、shiny 等包）、GeoFlow 和 Google Chart API 等，了解它们的优缺点及适用场景。

（5）认识可视化分析在大数据分析中的重要作用，理解其如何结合人类的感知认知能力和计算机的分析计算能力，以更直观、高效的方式洞悉大数据背后的信息、知识与智慧。

5.1 可视化技术简介

传统数据可视化是关于数据视觉表现形式的科学技术研究。其中，这种数据的视觉表现形式被定义为一种以某种概要形式抽提出来的信息，包括相应信息单位的各种属性和变量。数据可视化是一个处于不断演变之中的概念，其边界在不断扩大。其主要指利用图形、图像处理、计算机视觉及用户界面，通过表达、建模以及对立体、表面、属性和动画的显示，对数据加以可视化解释。与立体建模之类的特殊技术方法相比，数据可视化涵盖的技术方法要广泛得多。

传统数据可视化与信息图形、信息可视化、科学可视化以及统计图形密切相关。当前，在研究、教学和开发领域，传统的大数据可视化是一个极为活跃而又关键的领域。数据可视化实现了成熟的科学可视化领域与较年轻的信息可视化领域的统一。随着城市、交通、气象等数据容量和复杂性的与日俱增，传统数据可视化已经难以满足需求。在此背景下，大数据可视化应运而生，成为人类对信息的一种新的阅读和理解方式。在 GIS 领域，通过大数据可视化手段进行数据分析，可以从密密麻麻、错综复杂的数据中挖掘信息，再通过可视化的方式展示出来，使读者对数据的空间分布模式、趋势、相关性和统计信息一目了然，而这些在其他呈现方式下可能难以发现。随着大数据的发展，大数据可视化作为大数据处理的最后一个环节，将大数据开发工程师处理的数据进行再次整理，并将处理结果通过图表的形式呈现出来，使用户更加直观地看到数据的变化及趋势，为后期业务调整提供支持。

1. 传统数据可视化的几个基本概念

数据可视化的基本思想是将数据库中的每一个数据项作为单个图元元素表示，大量的数据集构成数据图像，同时以多维数据的形式表示数据的各个属性值。用户可以从不同维度观察数据，对数据进行更深入的观察和分析。数据可视化有以下几个基本概念。

（1）数据空间：由 n 维属性和 m 个元素组成的数据集构成的多维信息空间。

（2）数据开发：利用一定的算法和工具对数据进行定量的推演和计算。

（3）数据分析：指对多维数据进行切片、块、旋转等动作剖析数据，从而能多角度、多侧面地观察数据。

2. 传统数据可视化的理论模型

传统数据可视化的理论模型如图 5-1 所示。

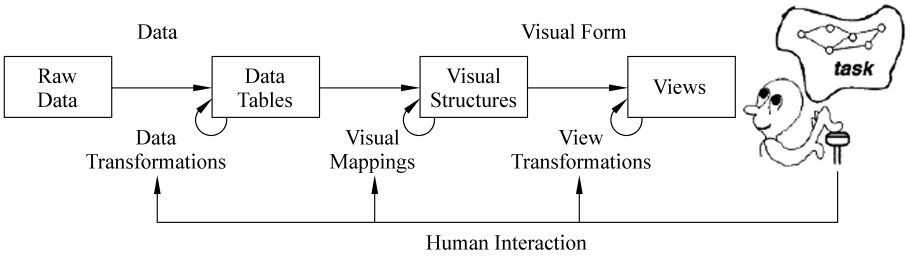

图 5-1　传统数据可视化的理论模型

其中，数据变换将原始数据转换为数据表形式（数据规范化）；可视化映射将数据表映射为可视化结构，由空间基、标记以及标记的图形属性等可视化表征组成（构建可视化结构）；视图变换则将可视化结构根据位置、比例、大小等参数设置显示在输出设备上（可视化输出）。

3. 传统数据可视化与大数据可视化对比

传统数据可视化和大数据可视化在多个方面存在显著区别，主要体现在数据规模、处理方式、实时性要求、交互性以及可视化目的等方面。以下是两者的主要对比。

（1）数据规模与复杂度：传统数据可视化通常处理的是较小的数据集，数据量和维度较少，主要用于描述和展示单一数据源或较小规模的数据。而大数据可视化处理的是大规模、多维度、高复杂度的数据集，数据量可能达到 TB 级别甚至 PB 级别。大数据可视化不仅关注数据的展示，还涉及从海量数据中提取有意义的信息，通常需要强大的分布式存储和计算平台来支撑。

（2）数据处理与存储：传统数据可视化的数据量较小，可以存储在本地数据库或单机系统中，数据处理和存储需求较低，处理的复杂度也相对较小。由于数据量庞大，大数据可视化通常需要使用分布式存储系统（如 HDFS）和分布式计算框架（如 MapReduce、Spark 等）。数据处理涉及大规模的聚合、筛选、清洗等操作，需要高效支持高并发的数据读取与展示。

（3）实时性与交互性：传统数据可视化大多为静态或周期性更新的数据可视化，用户

的交互性较低，通常用于呈现历史数据或某一时段的分析。大数据可视化更强调实时性和动态交互，能够处理实时流数据，并展示实时分析结果。支持复杂的交互功能，如数据过滤、钻取、聚合等，使用户能够灵活探索数据，获得更深层次的洞察。

(4) 可视化目的与深度：传统数据可视化的主要目标是展示数据的整体趋势、分类信息或基础统计结果，侧重数据的易读性和美观性，帮助用户快速理解数据。大数据可视化不仅关注数据的趋势展示，重点是可以从大量复杂的数据中提取潜在的有价值信息，帮助用户深入分析数据背后的模式、关系和预测未来。常常需要支持多维度分析、时序分析等复杂操作。

(5) 呈现方式：传统数据可视化主要使用简单的图表和表格，如条形图、折线图、饼图等，呈现方式直观且易于理解，适合展示小规模数据。大数据可视化为了应对大规模和高维度数据的复杂性，采用更为高级的可视化方式，如热力图、树状图、关系图、3D 图表、地理信息可视化、流图等。这些方式能够更好地呈现数据之间的复杂关系和多维数据。

总的来说，传统数据可视化主要适用于处理小规模、静态的数据集，而大数据可视化则应对海量数据、动态变化和复杂的数据分析需求，具有更高的技术要求和分析深度。

4. 大数据可视化分析方法

(1) 原位交互分析技术：进行可视化分析时，将在内存中的数据尽可能多地进行分析，称为原位交互分析。

(2) 数据存储技术：大数据是云计算的延伸，云服务及其应用深刻影响了超大规模数据库与存储社区。

(3) 并行计算：并行处理可以有效地减少可视计算占用的时间，从而实现数据分析的实时交互。

(4) 可视化分析算法：大数据的可视化算法不仅要考虑数据规模，而且要考虑视觉感知的高效法，需要引入创新的视觉表现方法和用户交互手段。

(5) 用户界面与交互设计：主要包括用户驱动的数据简化、可扩展性与多级层次、异构数据融合、交互查询中的数据概要与分流、表示证据和不确定性、时变特征分析、设计与工程开发等。

大数据可视化在发展过程中衍生出了分支：科学可视化——利用计算机图形学来创建视觉图像，帮助人们理解科学技术概念或结果的那些错综复杂而又往往规模庞大的数字表现形式。

5. 大数据可视化的重要性

从人类大脑处理信息的方式看，使用图形图表观察大量复杂数据要比查看电子表格或报表更容易理解。大数据可视化就是这样一种以最普通的方式向人类快速、简单传达信息的技术。通过大数据可视化能够有效地利用数据，帮助人们给诸如以下问题快速提供答案。

(1) 需要注意的问题或改进的方向。

(2) 影响客户行为的因素。

(3) 确定商品放置的位置等。

通过增加大数据可视化的使用，企业能够更快地发现所要追求的价值。创建更多的信

息图表，人们能更快地使用更多的资源，获得更多隐含的关系。人们可以意识到许多已知的信息，从而增加发现关键问题的可能性。它在看似没有任何联系的数据点之间创建连接，使人们能够分辨有用和没用的数据，最大限度地提高生产力，让信息价值最大化。

6. 大数据可视化的用途

（1）快速理解信息：通过使用业务信息的图形化表示，企业可以以一种清晰的、与业务联系更加紧密的方式查看大量数据，以制定决策。相对电子表格的数据分析，图形化格式的数据分析要更快，因此企业可以更加及时地发现问题、解决问题。

（2）标识关系和模式：即使面对大量错综复杂的数据，图形化表示也使数据变得可以理解。企业能够识别高度关联、互相影响的多个因素。这些关系有些是显而易见的，有些则不易发现。识别这些关系可以帮助组织聚焦于最有可能影响其重要目标的领域。

（3）确定新兴趋势：使用数据可视化，可以辅助企业发现业务或市场趋势，准确定位超越竞争对手的优势，最终影响其经营效益。企业更容易发现影响产品销量和客户购买行为的异常数据，并把小问题消灭于萌芽之中。

（4）方便沟通交流：一旦从可视化分析中对业务有了更新更深入的了解，就需要在组织间沟通这些情况。使用图表、图形或其他有效的大数据可视化表示在沟通中是非常重要的，因为这种表示更能吸引人的注意，并能快速获得彼此的信息。

7. 实施大数据可视化需要考虑的问题

实施一个新技术需要采取一些有效步骤。除了扎实地掌握数据外，还需要理解目标、需求和受众。准备实施大数据可视化技术时，先要做好以下准备工作。

（1）明确试图可视化的数据，包括数据量和基数（一列数据中不同值的个数）。

（2）确定需要可视化和传达的信息种类，如事务明细、累积聚合、比值比例等。

（3）了解数据的受众，并领会他们如何处理可视化信息。

（4）使用一种对受众来说最优、最简的可视化方案传达信息。

明确了数据属性和作为信息消费者的受众的相关问题后，就需要准备与大量数据打交道了。大数据给可视化带来新的挑战，4V（Volume、Velocity、Variety、Veracity）特性是必须要考虑的问题，而且数据产生的速度经常会比其被管理和分析的速度快。需要可视化的列的基数也是应该重点考虑的因素，高基数意味着该列有大量不同值（如身份证号），低基数则说明该列有大量重复值（如性别）。计算机诞生之前，科学的可视化行为就存在，如等高线图、磁力线图、天象图等。利用计算机的强大运算能力，人类可以使用三维或四维的方式表现液体流型、分子动力学的复杂科学模型。例如利用经验数据，科学可视化在天体物理学（模拟宇宙爆炸等）、地理学（模拟温室效应）、气象学（龙卷风或大气平流）模拟人类肉眼无法观察或记录的自然现象；利用医学数据研究和诊断人体；或者在建筑领域、城市规划领域或高端工业产品的研发过程中发挥重大作用。

虽然科学可视化的表现形式对于普通人比较陌生，但实际上其成果已经渗透到人们生活的每个角落。20 世纪 90 年代初期，信息可视化进入人们的视野，用于进行对异质性数据“抽象”部分的分析，帮助人们观察和理解抽象概念，扩大了人类的认知能力。科学可视化和信息可视化的差别比较微妙，因为科学可视化的大部分处理对象都是抽象的概念，在手段和

技术上也有大量共同之处，所以边界比较模糊。国外许多大型企业、科研机构都有相关部门专门进行数据可视化研究。媒体和政府机构也会对自己掌握的数据进行可视化分析，如犯罪地图等应用。在互联网上，那些掌握了大量用户活动信息、用户关系网或语料库的网站，如 digg、friendfeed、flickr 或大型电子商务网站等，都有实验性的可视化项目。

大数据可视化的开发和大部分项目开发一样，也是根据需求来筛选数据维度或属性，根据目的和用户群选用表现方式。同一份数据可以可视化成多种看起来截然不同的形式。有的可视化目标是为了观测、跟踪数据，所以要强调实时性、变化、运算能力，可能会生成一份不停变化、可读性强的图表。有的是为了分析数据，所以要强调数据的呈现度，可能会生成一份可以检索、交互式的图表。有的是为了发现数据之间的潜在关联，可能会生成分布式的多维的图表。有的是为了帮助普通用户或商业用户快速理解数据的含义或变化，会利用漂亮的颜色、动画创建生动、明了、具有吸引力的图表。还有的图表可以用于教育、宣传或政治，被制作成海报、课件，它们常出现在街头、广告、杂志和集会上。这类图表拥有强大的说服力，使用强烈的对比、置换等手段，可以创造出极具冲击力、直指人心的图像。国外许多媒体会雇用设计师，根据新闻主题或数据来创建可视化图表，对新闻主题进行辅助宣传。

5.2 数据可视化技术

5.2.1 文本数据可视化

文字是传递信息最常用的载体。在这个信息爆炸的时代，人们接收信息的速度已经小于信息产生的速度，尤其是文本信息。当大段的文字摆在面前，人们已经很少有耐心去认真把它读完，更多会先找文中的图片来看。这一方面说明人们对图形的接受程度比枯燥的文字要高很多，另一方面说明人们急需一种更高效的信息接收方式，文本可视化正是解药良方。教材里的解释图、自己笔记里总结的知识结构图，一直到现在经常用的思维导图等，其实都是简单、实用的文本可视化效果。本节将简单介绍文本可视化的基础概念，然后通过各类文本可视化的案例来阐述可视化之美。

文本数据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，是人们理解复杂的文本内容、结构和内在规律等信息的有效手段。海量信息使人们处理和理解的难度增大，传统的文本分析技术提取的信息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。将文本中复杂的、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，人们就能够利用与生俱来的视觉感知并行化处理能力快速获取大数据中蕴含的关键信息。

文本数据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、推理的能力得到充分发挥。文本数据可视化有效结合了机器智能和人类视觉，为人们更好地理解文本和发现知识提供了新的有效途径。文本数据可视化主要有以下作用。

（1）理解：理解主旨。

（2）组织：组织、分类信息。

（3）比较：对比文档信息。

（4）关联：关联文本的 pattern 和其他信息。

1. 文本数据可视化的流程

图 5-2 展示了文本数据可视化的流程。一般把对文本的理解需求分成 3 级：词汇级（Lexical Level）、语法级（Syntactic Level）和语义级（Semantic Level）。不同级的信息挖掘方法也不同，词汇级当然是用各类分词算法，语法级用一些句法分析算法，语义级用主题抽取算法。以上这些都在图 5-2 所示的文本信息挖掘中进行，其中文本预处理是过滤无效数据，提取有效词等；文本特征的抽取是指提取文本的关键词、词频分布、语法级的实体信息、语义级的主题等；文本特征的度量是指对多种环境或多个数据源抽取的文本特征进行深层分析，如相似性、文本聚类等。

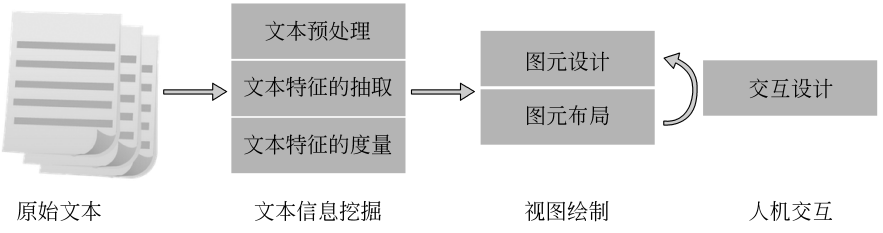

图 5-2　文本数据可视化流程图

2. 文本数据可视化类型

文本数据大致可分为 3 种：单文本、文档集合和时序文本数据。根据这些类型，文本数据可视化也可分为 3 类。

（1）文本内容可视化：文本内容可视化的方法主要基于关键词进行展示。标签云（Tag Cloud）是最常见的文本数据可视化方式，它通过不同大小和颜色的字体展示文本中出现频率较高的关键词，直观反映出重要的主题或概念。基于关键词的文本可视化包括文档散点图（DocuBurst），它通过径向布局展示词汇的语义层级关系。外层词汇是内层词汇的下义词，颜色的深浅反映了词频的高低。如图 5-3 所示。

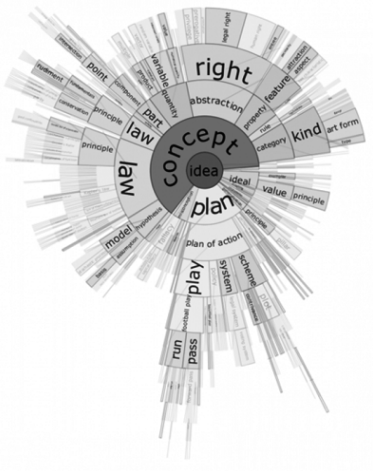

图 5-3　文档散点图

文档卡片（Document Cards）则是结合了文档中的关键词和关键图片进行可视化，布局在一张小卡片中。其中的关键图片是指采用智能算法抽取并根据颜色分类后的代表性图片。

（2）时序文本内容可视化：时序数据是指具有时间或顺序特性的文本，例如一篇小说故事情节的变化，或一个新闻事件随时间的演化。

① SparkClouds 在标签云的基础上，在每个词下面

增加了一条折线图，用以显示该词的词频随时间的演变，如图 5-4 所示。

图 5-4 SparkClouds 词云图

② 主题河流(ThemeRiver)是一种经典的时序文本可视化方法。光阴似水，用河流来隐喻时间的变化，几乎所有人都能非常好地理解。横轴表示时间，每一条不同颜色的线条可视作一条河流，而每条河流则表示一个主题，河流的宽度代表其在当前时间点上的一个度量(如主题的强度)。这样既可以在宏观上看出多个主题的发展变化，又能看出在特定时间点上主题的分布，如图 5-5 所示。

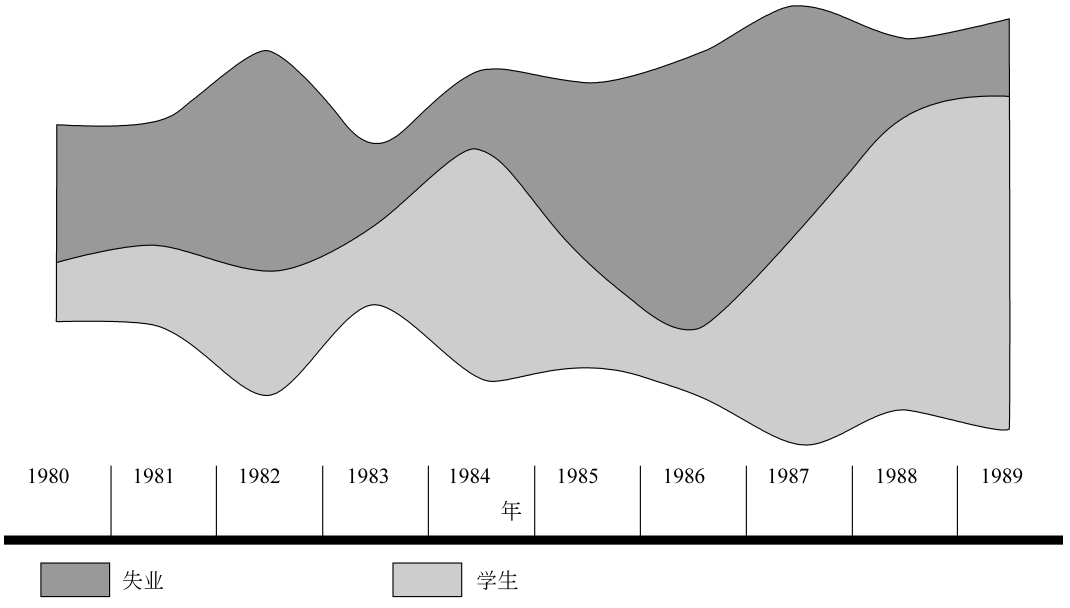

图 5-5 主题河流(ThemeRiver)

③ TIARA 结合了标签云，通过主题分析技术(Latent Dirichlet Allocation，LDA)将文本关键词根据时间点放置在每条色带上，并用词的大小来表示关键词在该时刻出现的频率。

因此，用 TIARA 就可以帮助用户快速分析文本的具体内容随时间变化的规律，而不是仅仅一个度量带变化。

④ TextFlow 也是 ThemeRiver 的一种拓展，它不仅表达了主题的变化，还表达了各个主题随着时间的分裂与合并。例如某个主题在某个时间分成了两个主题，或多个主题在某个时间合并成了一个主题，如图 5-6 所示。

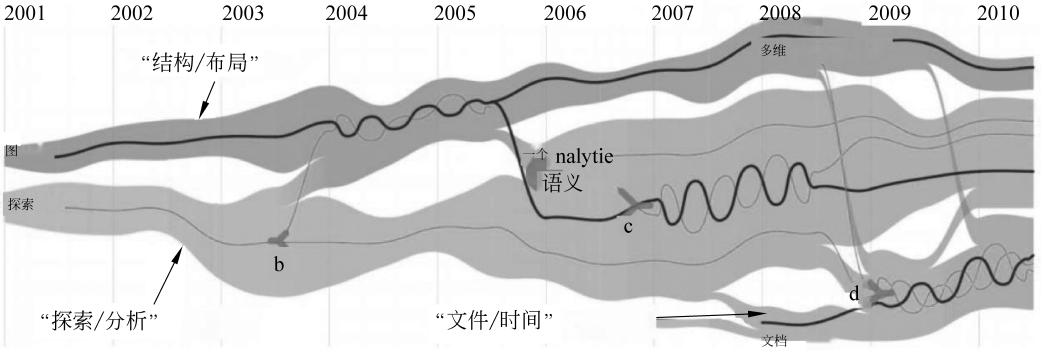

图 5-6　TextFlow

（3）文本特征分布模式可视化。

① TextArc 用来可视化一个文档中的词频和词的分布情况。整个文档用一条螺线表示，文档的句子按文字的组织顺序布局在螺线上，螺线包围着的是文档中出现的单词，每个单词的位置由其在文本中的频率和出现位置决定，饱和度用来映射词频。所以全局出现频率越高的词越靠近中心，而局部出现频率越高的词越靠近其相应的螺线区域。选中某个单词后，自动用射线关联到它在文中出现的位置，如图 5-7 所示。

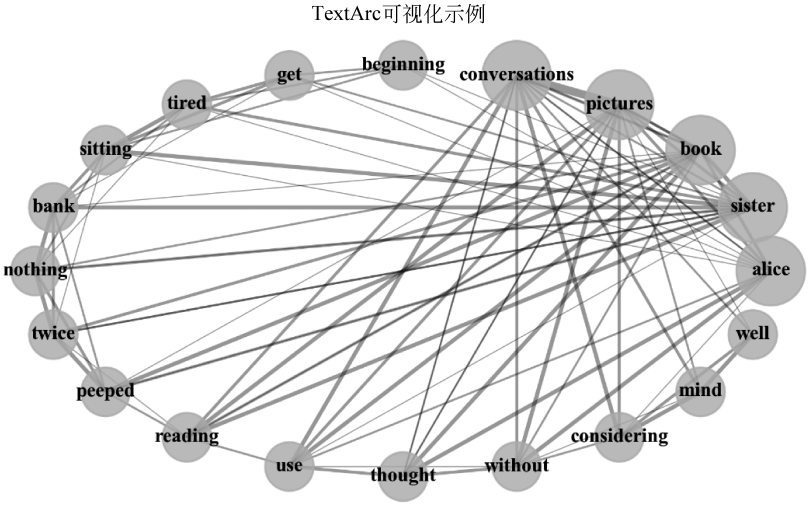

图 5-7　TextArc

② 文献指纹（Literature Fingerprinting）是体现全文特征分布的一项工作。一个像素块代表一段文本，一组像素块代表一本书。图 5-8 采用颜色映射句子长度，展示了 Jack London 和 Mark Twain 两个人的几部作品的可视化效果，从图中明显看出两人的写作风格迥异。

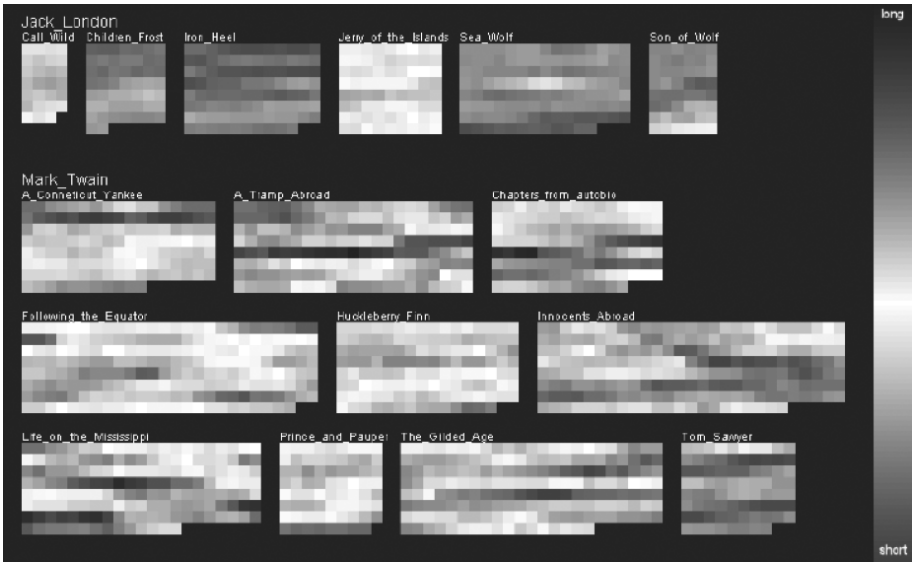

图 5-8 文献指纹(Literature Fingerprinting)

3. 情感分析可视化

情感分析是指从文本中挖掘出心情、喜好、感觉等主观信息。现在人们把各类社交网络当作感情、观点的出口,所以分析这类文本就能掌握人们对于一个事件的观点或情感的发展。图 5-9 是基于矩阵视图的客户反馈信息的可视化工作,其中的行是指文本(用户观点)的载体,列是用户的评价,颜色表达的是用户评价的倾向程度(暖色代表消极,冷色代表积极),每个方格内的小格子代表用户评价的人数,评价人数越多,小格子越大。

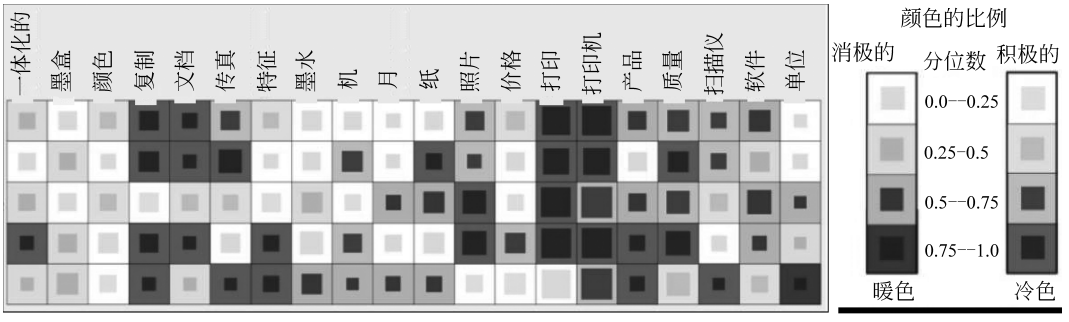

图 5-9 情感分析可视化

4. 文本关系可视化

顾名思义,文本关系可视化研究的是文本或文档集合中的关系信息,比如文本的相似性、互相引用的情况、链接等。说到关系布局,一般都是使用树或图。

(1) 文本内容关系可视化。

① 单词树(Word Tree)是把文本中的句子按树状结构布局,可以很好地看出一个单词在文本中出现的频率和单词前后的联系,如图 5-10 所示。

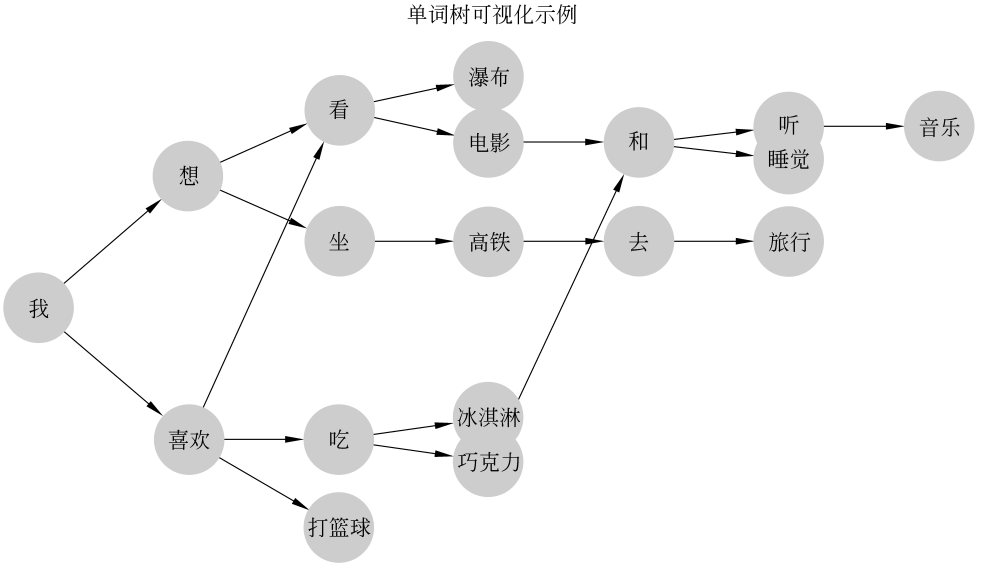

图 5-10　单词树(Word Tree)

② 短语网络(Phrase Nets)是经典的力导向图结构，图中的节点是从文本中挖掘出的词汇级或语法级的语义单元，边代表语义单元的联系，边的方向即短语的方向，边的宽度是短语在文本中出现的频率，如图 5-11 所示。

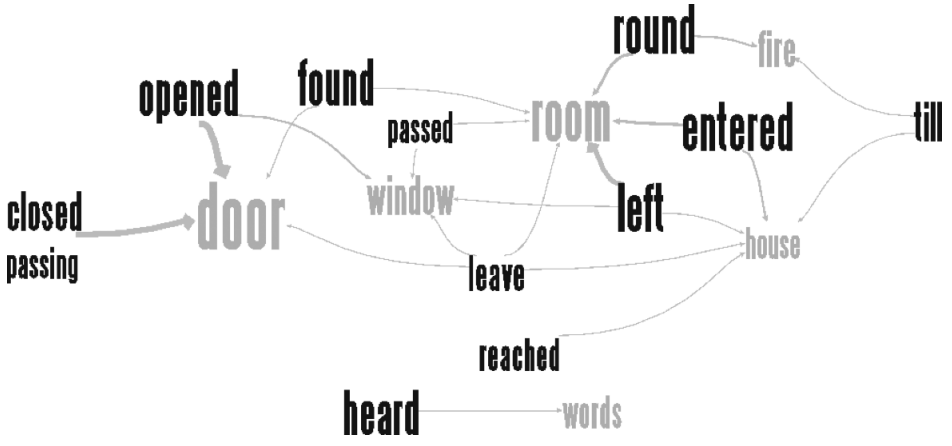

图 5-11　短语网络(Phrase Nets)

(2) 文档集合关系可视化：文档数量到一定量时，再针对文本做可视化，就会过于复杂，所以通常可以对单个文档定义一个特征向量，利用向量空间模型计算文档间的相似性，并采用相应的投影技术呈现文档集合的关系。

① 星系图(Galaxy View)把一篇文档比作一颗星星，用投影的方法把所有文档按照其主题的相似性投影为二维平面的点集，星星离得越近，代表文档越相似，因此可以从一个星系图非常直观地看出文档主题的紧凑和离散程度，如图 5-12 所示。

② 主题地貌(Theme Scape)是对星系图的改进。把等高线加入投影的二维平面中，文档相似性相同的放在一个等高线内，再用颜色来编码文本分布的密集程度，把二维平面背景