文档内知识挖掘与服务研究

陈　静/著

科学出版社

本书受国家自然科学基金青年项目“图书层次主题自动标引研究”（项目编号：71303089）资助

文档内知识挖掘与服务研究

陈　静/著

科学出版社

北京

内 容 简 介

随着互联网应用的发展，电子文档信息资源已成为用户在大数据环境下获取知识的重要信息源。然而，现有文档知识挖掘与服务研究的粗粒度现状与信息用户需求的精细化趋势之间的矛盾日趋严重：现有文档知识挖掘与服务研究尚未深入文档内部多粒度的主题信息，无法彰显文档内部知识，严重阻碍了信息资源管理与应用的发展。基于此，本书以文档内多层结构知识组织为切入点，以层次主题挖掘为主要手段，构建文档内多粒度知识挖掘的模型与方法体系，进而结合认知与行为科学理论方法，从不同用户理论维度视角，研究构建文档内知识服务利用及用户行为的理论体系，以实现文档内多粒度知识挖掘与知识服务，拓展文档信息资源研究内容，推进文档信息资源管理与应用发展。

本书可作为信息管理、情报学与计算机应用等专业人员的科研与教学参考用书。

图书在版编目（CIP）数据

文档内知识挖掘与服务研究 / 陈静著. —北京：科学出版社，2018.7

ISBN 978-7-03-057366-7

Ⅰ. ①文… Ⅱ. ①陈… Ⅲ. ①信息管理-研究 Ⅳ. ①G203

中国版本图书馆 CIP 数据核字（2018）第 095476 号

责任编辑：邓 娴 / 责任校对：王晓茜

责任印制：霍 兵 / 封面设计：无极书装

科 学 出 版 社 出版

北京东黄城根北街 16 号

邮政编码：100717

http：//www.sciencep.com

新科印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2018 年 7 月第 一 版 开本：720 × 1000 1/16

2018 年 7 月第一次印刷 印张：12

字数：242 000

定价：86.00 元

（如有印装质量问题，我社负责调换）

安徽大学图书馆 藏书 www.lib.ahu.edu.cn

前　言

近年来，随着电子文档信息资源迅猛增长，电子文档信息已成为用户在现代环境下获取知识的重要信息源；同时，处于知识爆炸时代的文档用户，越来越需要方便准确地查找和快速有效地学习大量电子文档内部的各种新知识，这就需要计算机系统自动深入文档内部进行知识挖掘，进而向其提供细粒度的知识服务。然而，现有文档知识挖掘与服务研究很少深入文档内部，无法彰显文档内的多粒度知识，严重阻碍了网络与数字图书馆中高价值电子文档信息资源的管理及应用发展。

本书从电子文档粗粒度服务问题出发，针对文档内挖掘与服务的研究不足，着力于构建文档内知识挖掘与服务的理论模型与方法体系，并系统研究其相关用户行为，以满足现代信息资源管理理论研究与应用发展的时代需求。具体来说，本书面向文档内多层结构知识组织，以目次解析为基础，以层次主题挖掘为主要手段，研究文档内多粒度知识挖掘的模型与方法，进而研究构建文档内知识服务利用及用户行为的理论体系，以实现文档内多粒度知识挖掘与知识服务，拓展文档信息资源研究内容，推进文档信息资源管理与应用发展。

本书首先阐述文档内知识挖掘与服务的理论基础和基本模型，对现有文档内知识挖掘与服务相关研究进行系统的分析和评述，找出文档内知识挖掘与服务研究的关键问题，为本书奠定研究基调与理论框架，是本书的基础所在；其次以文档内知识挖掘与服务的几种主要资源类型为对象，紧扣文档内知识挖掘与服务研究主题，通过图书目次知识获取与利用、图书文本主题挖掘与层次组织、长文档内知识挖掘与服务等章节，以清晰的资源脉络构建出文档内知识挖掘与服务的基础模型和方法体系；再次以文档内知识挖掘与服务的用户为对象，从用户使用文档内知识服务的行为、绩效与体验等关键问题出发，通过文档内知识服务的用户行为数据采集、文档内知识服务的可用性、文档内知识服务的用户认知等章节，探讨并建立较为完善的文档内知识服务利用及用户行为的理论体系，资源视角与用户视角互为补充，构建出文档内知识挖掘与服务的理论模型和方法体系；最后为本书总结与展望。

本书突破现有文档知识挖掘与服务研究的粗粒度局限性，将文档知识挖掘与服务研究深化到文档内多层结构知识的细粒度层面，由此产生一系列新的基础理论模型与科学方法，其学术观点与理论体系具有基础性及原创性的特征，理论创新程度高。

本书是全面和深入研究文档内细粒度知识挖掘与服务的著作。本书深入文档内部，从文档内部语义与结构等多角度出发，将文档主题分析、文档知识发现、文档知识服务、用户认知与行为、人机交互等理论、模型、方法结合，从文档内知识挖掘、知识间的语义关联组织、知识服务的用户行为与认知分析等多个维度，系统全面地研究文档内知识挖掘与知识服务，建立文档信息资源细粒度管理与利用的基础理论和方法体系。本书的研究成果具有较高的学术价值，对深入文档内部进行知识挖掘、知识组织与知识服务研究具有重要的指导价值，对文档信息资源的管理与利用及文档用户研究有重要的推动作用，对推动和促进信息管理学科及其他相关学科的研究发展有重要启发意义。

本书前沿性和现实性强，集信息管理与计算机应用专业之所长，既具有理论系统性和创新性，又具有全面而深入的实践性。由于作者水平有限，书中难免有不足之处，希望广大读者给予批评指正。

陈　静

2018 年 4 月 16 日

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

第 1 章　文档内分析基础

处于知识爆炸时代的文档用户，越来越需要方便准确地查找和快速有效地学习大量电子文档内部的各种新知识，这就需要计算机系统自动深入文档内部进行知识挖掘，进而向用户提供多粒度的知识服务。知识挖掘是从数据集中识别出有效、新颖、潜在有用以及最终可理解的模式的过程；而知识服务则是通过知识挖掘，从显性和隐性的知识资源中提取用户所需，用以解决用户问题的信息服务过程。随着数字文档资源的迅速增长以及开放获取，准确地从文档中找到需要的信息以及快速地理解文档等需求显得越来越迫切。然而，现有文档知识挖掘与服务研究很少深入到文档内部，无法彰显文档内的多粒度知识，严重阻碍了网络与数字图书馆中高价值电子文档信息资源的管理和应用发展。文档内分析（within-document analysis）研究结合文本分析技术和数据可视化技术，可以有效地帮助人们快速理解文档内部的结构、内容和规律。近年来关于文档内分析方面的研究层出不穷，为了让读者了解文档内分析，本章在概述文档内分析的基础研究和可视化分析工具的基础上，考虑到由于文档内知识服务需要精准分析用户认知行为，本书多个章节将涉及有关眼动、认知方面的内容，本章还对认知负荷测量及其在情报学中的应用进行梳理和总结。

1.1　文档内分析理论与方法

1.1.1　文档内分析功能

针对用户的不同阅读任务与需求，现有文档内分析的功能主要集中在文档概览、文档内导航与浏览、文档内检索三个方面。

1. 文档概览

通过文档概览，用户不需查看全文就能够了解文档的主要内容、结构和内在规律。文档概览几乎是所有文档内分析工具所具有的功能，标签云就是一种典型的文档概览形式，通过对文档中高频词汇的展示能让用户大致理解文档的主要内容；Document Card（Strobelt et al.，2009）在标签云的基础上通过自动提取文档中重要的图片和文字，将其综合到一系列连续的卡片上，用户通过这些卡片就能

快速地理解文档的关键内容；TextArc 把一篇文档中所有出现的字符行可视化在椭圆圆周上，从而展现了文档中词的词频、位置及共现关系，用户通过交互操作与概览图形进行互动，能快速发现文档中的词汇关系与分布，了解文档概况。此外，还有通过语义关系概览让用户快速了解文档内在规律，通过主题结构概览让用户快速了解文档主题等概览形式。

2. 文档内导航与浏览

文档概览能帮助用户对文档全貌有一定的了解，在文档概览形式基础上，文档内导航与浏览能为用户查看文档详情提供导航，用户可以与文档概览进行互动，选择性地浏览文档的相关详细信息。Popout Prism（Suh et al.，2002）是一个“Overview + detail”文档界面，它可以帮助用户顺利地转换完整文档的概览，进而迅速找到关注的详细信息；Overview Scrollbar（Mizoguchi et al.，2013）使用滚动条显示一个完整文档的概览，进而通过点击概览任意一处能快速跳到文档中相应的地方；DocuBurst（Collins et al.，2009）以径向空间填充图形形式展示文档中关键词汇之间的层次关系，当用户选择图中一个节点时，对应词汇的分布情况在线性的以长方形表示的文本片段可视化图形“瓷砖”中显示，通过选择任意一个瓷砖就能浏览相应的文档片段中的词汇详细分布情况。

3. 文档内检索

在集中进行多文档检索时，检索到的文档一般会通过文档代理以列表的形式提供给用户，然后用户可以通过文档代理来判断文档相关性并且选择打开可能相关的文档进行查看。但是，用户可能还希望找到文档中与查询相关的具体部分，文档内检索功能恰好能满足这一需求。文档内检索不仅能找到查询词在文档中出现的所有地方，同时，文档内检索还可以提供查询结果概览。例如，TileBars（Hearst，1995）可以在多文档检索中对每个文档进行查询词的可视化检索；SmartSkim（Harper et al.，2002）支持一个文档内的可视化检索。

1.1.2　文档内分析策略

现有文档内分析策略主要有词汇、内部语义关系及主题三种分析策略。

1. 词汇分析策略

词汇分析的对象主要是基于文档中的词汇，它通过词汇的不同呈现方式来展示文档的内容特征，主要包括词频统计以及词汇分布两种策略。

1）词频统计

词频统计是一种词汇分析研究方法，它通过统计一定长度范围内语料或文本中每个词出现的次数和频率，分析统计结果，以便描绘词汇规律（刘洪波，1991）。词频统计有两种计算方法：一种是基于统计的方法，需要对整篇文章进行分词进而对每个单词出现的频率进行统计；另一种是基于匹配的方法，即利用字典对文档进行扫描及匹配，以统计出所有单词出现的频率。通常情况下，以出现频率较高的词来表示整个文档，可以对文档内容进行概述。典型的代表就是 Wordle（Viegas et al.，2009），它通过统计方法快速分析文本或网站的词频，且支持文字字体选择和用户自定义颜色，以多种风格展示文本词汇特征；ManiWordle（Koh et al.，2010）在 Wordle 的基础上进行改进，支持用户对全局以及单个词汇的字体、颜色、构图进行自定义操作；DocuBurst（Collins et al.，2009）进一步考虑词汇之间的语义关系，将词频与人造词汇数据库 WordNet（Fellbaum and Miller，1999）进行结合，以径向空间填充圆形形式展示文本结构，以词汇在 WordNet 中的上下位关系和词频来反映文档语义内容。

2）词汇分布

在信息检索与服务领域，用户检索到相关文档后，还需在文档中进一步确定相关区域，文档上下文和逻辑结构在这类研究中占据着重要的作用，词汇分布是文档内检索的重要体现。FindSkim（Harper et al.，2004）高亮显示文档中所有查询词的变体，文档界面会定位到查询词首次出现的地方；TileBars（Hearst，1995）强调词汇分布信息对于文档相关性判断的重要性，允许用户查看检索文档的相对长度、对应文档中查询词的频率以及分布情况，来帮助用户进行相关性判断；Scrollbar（Byrd，1999）在此基础上进行改进，考虑了查询词的权重以及以更直观的滚动条方式展示了查询词在全文中的分布情况，来帮助用户快速定位到文档中最相关的部分。应用 TileBars 的思想，Schwartz 等（2010）引入 Focus + Context 模型将文档内搜索结果进行可视化展示，即做成全文中查询词分布连续直方图，通过直方图识别特定的信息，同时考虑查询词的上下文环境。

2. 内部语义关系分析策略

对自然语言文本中的文本单元（词、句子或段落）间的语义关系进行全局分析，能帮助用户深入理解整个文档的内部结构和语义关系，快速抓住文档内部规律。NLPwin（Leskovec et al.，2004）对每个句子抽取主谓宾三元组逻辑结构，应用跨句指代处理、共引处理、语义规范来完善三元组关系并将它们合并成语义图，以反映文档中核心词及其语义关系；Phrase Nets（Ham et al.，2009）根据用户指定的某一种语义关系，通过句法结构和基于文本模式匹配分析，以网络图的形式来展示词汇和词汇之间的关系，同样的文档以不同的语义关系可

以展示出不一样的视角，帮助用户理解非结构文档中关键概念及其之间的关系。命名实体的语义关系是众多文档内分析的方向，如 Contexter（Grobelnik and Mladenic，2004）将关键的人名、地名、术语等特定的命名实体抽取出来，以网络图的形式表现实体之间的共现关系，通过相关的关键词和其他命名实体来反映实体的背景信息；Fanlens（Lou et al.，2008）以径向空间填充方式动态地呈现了文档中命名实体的层次关系；Jigsaw（Stasko et al.，2008）是一个可视化分析系统，通过提供多种协调界面展示不同文档中命名实体之间的关系，帮助调查分析师更有效地检查文档和命名实体，从而做出决断。Chang 和 Collins（2013）使用词汇数据库抽取物理实体并计算其出现分数，并以 3D 图像形式展现了文档中固有的抽象和空间语义关系。此外，不少学者将语义关系以树状结构进行展示，Word Tree（Wattenberg and Viegas，2008）以节点表示词或句，以树状形式表示它们的层级关系，用字体的大小表示词或句出现的次数，能帮助用户查看查询词在文中出现的所有上下文语义关系；Netspeak（Potthast et al.，2011）同样以树状形式展示文本中常见的上下文结构，来帮助用户在写作时选择合适的词语；刘春江等（2011）结合了放射状布局和树布局的特点，将英文文本中高频词汇及词汇间的上下文关系以放射状树布局进行可视化展示，帮助人们快速理解英文文本的内容。

另外，还有使用本体来表现文档内的语义关系，本体是一种能在语义和知识层次上描述领域概念的建模工具，其目标是捕获相关领域的知识、确定该领域内共同认可的词汇。根据知识模型可以将本体分为表示普遍通用抽象概念的顶层本体、扩展顶层本体针对特定领域概念的领域本体和描述语言知识的词汇本体三种类型。薛中玉等（2009）开发了针对仪表领域文档的分析工具，通过建立仪表领域的专业词汇本体库，对仪表领域的文档进行分析，显示了文档中核心概念以及概念之间的关系，并以图形化方式展示分析结果；Nessah 和 Kazar（2012）应用词汇本体 WordNet 与领域本体，将文档生成结构化的元数据，通过概念图表现文档内概念及其关系的语义知识；iJADE InfoSeeker（Lim and Lee，2007）是一个帮助用户找到和分析中文网页文档并将文档内容以语义网形式展现的智能系统，其主要特点就是使用本体来分析中文文档语义并使用语义网来组织语义信息，以及使用本体来确定其主题。同时本体还被用来在不同的学术领域对科学文档进行注释，用来表示整个文档的结构、修辞元素和相关信息，丰富文档内容的表现形式，提高内容的可操作性和交互性以及文档之间的关联性，加深读者对文档内容的理解和满足读者个性化的阅读需求。

3. 主题分析策略

文档主题分析主要是通过揭示文档或文档集所包含的主题来帮助用户快速理

此为试读，需要完整PDF请访问：www.ertongbook.com

解文档。特别是对于长文档，用户可以通过主题分析快速判断其是否需要进一步阅读全文，既能节省时间也能降低用户的认知负担。

目前，基于文档主题的分析多是面向文档集进行的，一些研究通过词频统计来表示整个文档的主题信息，如 ThemeRiver（Havre et al.，2000）将文本数据进行标注统计，按照主题进行分割，以河流形式展示了主题在不同时间的分布情况以及随着时间变化的趋势；NewsLab（Ghoniem et al.，2007）、MemeTracker（Leskovec et al.，2009）和 Visual Backchannel（Dork et al.，2010）等在此基础上，将文本主题用于跟踪新闻、博客、Twitter 事件随时间的变化。为了抽取潜在的主题信息，有学者又利用主题模型和 ThemeRiver 技术来表示文本流，如 TIARA（Wei et al.，2010）系统通过主题模型 LDA（latent dirichlet allocation）（Blei et al.，2003）抽取文本主题将其展现在 ThemeRiver 河流中，而且展现了相应主题下的关键词来表示详细信息；TextFlow（Cui et al.，2011）以河流形式展现了主题随时间的变化以及展现了主题变化的关键点和相对应的关键词。因为线性的时间轴对于用户来说很难推断出不同时间点的语义关系，层次展示的方式应运而生，ThemeCrowds（Archambault et al.，2011）对 Twitter 中主题关键词以嵌套的长方形表示不同的主题层次关系，展示了主题随时间的变化关系；Smith 等（2014）使用分层潜在狄利克雷分配（hierarchical latent Dirichlet allocation，hLDA）（Blei et al.，2004）主题算法，以一个拥有各种色彩的阳光图来表示整个文档集的层次主题结构。

此外，由于单文档特别是图书这种长文档，相比短文档来说，拥有更加复杂的语义结构，往往涉及多个主题，主题也可以作为一个基本元素来对单文档内容进行分析，TOPIC ISLANDS（Miller et al.，1998）利用小波技术通过相关词频来抽取一系列主题，以层次岛屿进行可视化展示，用于表示文档中不同主题之间的关系；Nishihara 等（2009）提出一种使用亮色和阴影表示主题相关性的文本可视化方法，即亮色句子表示和指定主题相关，阴影句子表示和指定主题不相关，从而帮助用户找到和主题相关的部分，抓住文档中句子和主题的关系；Lu 和 Liu（2011）采用 LDA 算法进行文档主题分析，提出了一个多层主题地图，包括主题层、关键词层和资源层，帮助用户挖掘新的知识以及它们之间的联系。Wang 等（2013）提出一个针对长文档进行主题结构分析的模型，把文档分成多个片段，采用主题算法 LDA 进行主题抽取，以层次主题超图形式表示整个文档的主题结构，用户通过交互式操作可以分析主题演化、主题多样性以及主题交互；施乾坤（2013）将 LDA 模型针对语料集生成的主题词权重和 tf-idf 相结合来计算主题词组的权重，然后确定主题词布局，以标签云可视化形式表示单篇文档的主题。

1.1.3　文本分析方法

文档内分析主要是对文档中自然语言的词、句、段的处理，因此文档内分析

所使用的方法必然和自然语言处理的技术与方法相关。在文档内分析中使用较多的方法主要包括词袋模型、命名实体识别、模式匹配和主题模型等。

1. 词袋模型

词袋模型（bag of words，BOW）是一种用于文本的经典表示方式，最早用于文本分类、文本检索等领域，该模型基于这样的假设：一个文本，忽略其词序和语法、句法，将其仅仅看作一个词集合，或者说是词的一个组合，文本中每个词的出现都是独立的，不依赖于文本中其他词是否出现，或者说当这篇文章的作者在任意一个位置选择一个词汇都不受前面句子的影响。词袋模型对自然语言进行简化，能方便快捷地设计出来，被广泛地应用到文档分析处理中，Wordle（Viegas et al.，2009）、ManiWordle（Koh et al.，2010）通过对文档建立词袋模型，以文档中高频词对文档内容进行表示，在此基础上，加入时间元素，如 SparkClouds（Lee et al.，2010）、Tag River（Forbes et al.，2011）表现文档集中关键词随时间的变化情况，词袋模型也被应用在文档内检索中，如 TileBars（Hearst，1995）对检索文档建立词袋模型，通过检索词与文档中词汇进行匹配，来判断文档的相关性。虽然词袋模型假设从计算效率来看是有意义的，但其实是不切实际的，因为在许多语言建模应用中，如文本压缩、语音识别、预测文本输入等，词汇顺序是相当重要的（Wallach，2006），同时，两个具有相同词汇的句子，由于词汇顺序不同，表达的主题也会不同；对于整篇文档来说，也没有考虑词汇之间的逻辑关系以及词汇在文档中不同位置的权重问题。

2. 命名实体识别

命名实体识别是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。评定一个命名实体是否能被正确识别，主要体现在实体的边界是否正确以及实体的类型是否标注正确两个方面。命名实体识别方法主要可分为基于规则和词典的方法、基于统计的方法以及两者混合的方法。基于规则和词典的方法是早期使用的方法，是通过语言学专家手工构造规则，建立知识库和词典进行识别的方法，这种方法依赖于具体语言、特定领域和文本风格，知识库、词典编制耗时且可移植性差，具有一定局限性；基于统计的方法是利用人工标注的语料进行训练，不需要广博的语言学知识，而且可以在短时间内完成，这种方法较大地依赖于语料库，然而可以用来命名实体识别的大规模通用语料库又比较少。由于两种方法各有优劣，目前大多数情况下都是两种方法的结合使用。FacetAtlas（Cao et al.，2010）通过应用基于特定领域模型，例如，医学模型的命名实体识别依赖于标准命名实体识别和主题模型两种方法来抽取多方面的实体。前一种方法能容易地抽取组织、地点、时间等实体，后一种方法通过分析文档中

的主题线程，使用相关的关键词作为实体。Jigsaw（Stasko et al.，2008）使用了免费的商业实体识别系统进行自动实体识别，同时允许手动识别或调整实体的识别。

3. 模式匹配

在文档内关系分析中，需要识别自然语言中词汇、短语之间的语义关系，主要有三种识别方法。第一种，基于词共现的方式，但是这种方法往往分析不精确，会导致文本关系混乱，并且词共现关系往往只能显示一般性关系，在任何复杂的文档中都存在着更加特殊的关系，如人拥有东西、行动发生在某个地点等，这种方法就难以表现出这种复杂的语义关系；第二种，使用基于句子语法结构的模式，就是把一个句子分成关系连接的词对，词对之间的关系可以是几十种中的任何一种，如 A and B、A of B，这种方法虽然能避免第一种方法的缺点，但是这种方法分析起来比较复杂，耗时比较多；第三种，基于模式匹配的处理方法能做到更加高效准确，这种方法是根据具体的需求，从自然语言的语法中提取某一成分或几种成分的交融来作为模式，根据模式识别出自然语言中所需要的信息，而不必受制于语法（封春升和郝爱民，2006）。例如，定义（X，Y）表示文档中每一个“…X's Y…”，来找出文档中所有的这种从属关系。Phrase Nets（Ham et al.，2009）使用正则表达式定义了大量的潜在有趣的模式，找到文档中符合该模式的所有地方，并进行抽取，最终将语义关系以语义图的形式进行展示。

4. 主题模型

通过词袋模型、语义分析等使用词频来表示整体的主题信息，这些方法仅仅表示一个只包含一个复杂主题的文档部分，为了抽取连续关系的词汇的潜在主题信息，一系列主题概率模型如 LDA（Blei et al.，2003）、PLSI（probabilistic latent semantic indexing）（Hofmann，1999）和 TTMM 被提出。这些方法用来表示多个主题混合的文档。应用最广泛的就是 LDA，LDA 是一个无监督贝叶斯算法，使用词袋模型进行统计主题建模，将每个文档看成无序词汇的集合，每个文档被表示为一些主题上的一个概率分布，每个主题是词汇上的一个概率分布，LDA 具有模块化和可扩展性的优势，作为一个概率模块，LDA 很容易嵌入到一个更复杂的模型中，目前也有许多 LDA 的扩展性应用。由于 LDA 无法揭示主题之间的关系，hLDA（Blei et al.，2010）是 LDA 的层次结构变体，使用嵌套中国餐馆过程（nCRP）作为非参数先验来对 LDA 进行层次扩展，主题被安排在一个树状结构上，根节点表示一般主题，叶子节点表示更详细的主题，生成的层次主题模型灵活性强，能适应不断增长的数据集；TIARA（text insight via automated responsive analytics）（Wei et al.，2010）将 LDA 应用于文档内分析，把主题安排在时间序列上，展示了主题随时间的变化情况。

1.2　文档内可视化分析工具

比起浏览和理解文本信息，高度发达的视觉能力使人们能更快速地抓住一张图片的内容（Kamada and Kawai，1988）——人们能识别一张图片中元素的空间结构，并能很快注意到这些元素之间的关系。可视化技术结合计算机图形学、人机交互、认知科学等学科的理论与方法，以视觉符号的形式对文档中难以通过文字表达出来的复杂的内容、结构、内在规律进行展示，并向用户提供与这些视觉符号进行快速交互的功能，使人们的视觉认知、关联、推理等能力得到充分发挥，能帮助人们更好地理解文档中的内容、结构和内在规律。目前关于文档内分析的工具基本上是将文本分析得到的结果以可视化的形式展现出来，同时提供高亮、动态转换、向下钻取、关联更新、缩放、焦点加上下文等交互功能，以帮助用户有效地理解文档。

1.2.1　文档内可视化的基本模型

1. 标签云

标签云是关键词的视觉化描述，是一个用来可视化用户生成标签的有效介质。标签云大致分为两类：静态标签云和动态标签云。最典型的静态标签云是指标签是独立的词汇，词汇根据重要性以不同字体大小或颜色表示，标签按照字母顺序或重要性或频率排序，以长方形形式水平逐行布局。许多研究人员在这个方法基础上进行了改进，一些研究侧重于解决空白空间、重叠标签、特定界限的限制等常见问题。而动态标签云主要表现文本流的内容变化，SparkClouds（Lee et al.，2010）把标签云和折线图相结合，有效地表示主题随时间的频率变化趋势。

2. 网络图

网络图主要是将基于语义网络的概念用于文档内语义关系的可视化展示。语义网络是自然语言理解及认知科学领域研究中的一个概念，用来表达复杂的概念及其之间的相互关系，是一个有向图，其顶点表示概念，而边则表示这些概念间的语义关系，从而形成一个由节点和弧组成的语义网络描述图。例如，NLPwin（Leskovec et al.，2004）展现了文章 *Long valley volcano activities* 中核心词的语义关系，并以黄色灰色节点代表主语和宾语，以蓝色谓语标签表示这些节点之间的关系，每个节点通过一些属性描述，帮助理解节点内容。

3. 树图

树图是一种流行的利用包含关系表示层次化数据的可视化方法。层次化数据的可视化就是专门适用于呈现具有层次结构的数据的可视化技术，尤其强调对其中层次和包含关系的可视化呈现；采用不同的视觉符号来表现这种层次关系决定了层次化数据可视化的主要不同类别。其中一种就是节点链接图，数据中的个体通过二维或三维空间中的点、球或其他形式表示，个体之间的关系使用节点之间相连接的线或曲线表示。节点链接图可以表示任意图结构，当表示层次化数据时就成为树形结构，能够清楚地呈现节点间的层次关系。如 Word Tree（Wattenberg and Viegas，2008）中节点表示词或句，字体的大小表示词或句出现的次数。由于这种点线间的空白浪费了大量空间，学者提出了另外一种可视化方法就是树图，数据中的个体通过具有一定面积的块、体来表示，而个体之间的关系通过节点之间空间位置的包含关系来呈现。如 Treemap（Johnson and Shneiderman，1991）使用长方形表示节点，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，展示了文本的层次结构。这种方法虽然充分地利用了空白屏幕空间，但是节点之间的关系并没有树形结构清晰。

此外，学者对树图进行扩展提出了径向空间填充方式，使用了类似饼图的形式，以二维扇形或环形表示节点，沿着同心圆半径增加的方向以相邻的空间关系来表示数据的层次关系，这种方法更好地表现了层次结构，并广泛地应用在文本可视化中。如 DocuBurst（Collins et al.，2009）从中心根节点到外围遵循一条语义路径，根节点是通过搜索一个感兴趣的词或它的同义词来选定的，一旦根节点选定，可视化图形就以它的下位关系词来填充；点击一个节点，它的同义词集和整个子树被选择，能看到整个文档中的词共现模式，通过几何和语义缩放的交互，也能看到单个词汇在文档中的详细分布。

4. 相关性轮廓图

相关性轮廓图就是信息相关性的图形表示，目的是能让用户鉴定文档的相关部分，而通过与其交互就能直接移到文档的相关部分。一种方法是以一个图形展示每个词的位置权重，然后与图进行交互移到文档中对应的单词处。如 TextArc 把整个文档以行为单位，外圈列出了文档中的所有词，内圈列出了所有不重复的词，出现越多的词，会以越明亮的颜色显示，在文中分布越均匀的词，会出现在越接近圆心的位置；点击或检索一个单词，椭圆形中会显示射线分布，文本视图会显示使用这个词的所有行，并且词会高亮显示，同时也能查看共现关系的词。另外一种方法就是使用一个聚集类型图（如条形图）显示平均关联状态值，以单

词的位置范围作为条形图中的一条，如 Schwartz 等（2010）引入 Focus + Context 模型，将文档分段以条形图显示，查询词以不同颜色在条形图中分布显示，高度表明查询词出现频率，通过关联更新、向下钻取两次交互操作，得到更加细粒度的视图，找到文档中最相关的部分及上下文信息。

1.2.2　文档内可视化分析工具分类

1. 基于文档内词汇分布的可视化分析工具

通常，大致了解文档内容最快的方式就是浏览最频繁使用的词汇，了解词汇在文档中的分布情况也能从某种程度上帮助用户理解文档，如能揭示和特定词汇相关的文档区域。

1）ManiWordle

ManiWordle（Koh et al.，2010）是文字云在线生成器，是被广泛使用的一个有趣并且简单的工具，通过快速地分析文档的词频，并以多种风格显示高频词汇，帮助用户大概了解一篇文档的主要内容。将文档输入，就会生成文字云，用词的尺寸大小表示词频高低，用户可以通过设置参数自定义显示词汇数量、词汇布局、颜色风格、字体等，同时也可以对单个词汇的字体、颜色、构图进行调整，布局充分地利用了词汇间的空隙，可视化结果美观。

但是 ManiWordle 只是对高频词汇的简单罗列，并没有考虑词汇之间的语义关系，也没有考虑文档的逻辑结构。

2）TextArc

TextArc 能够在单个页面或屏幕上显示整个文档（主要是 email、新闻故事、学术论文等中等大小的原始文本）中词汇的频率和分布情况，帮助人们根据高频词汇快速感知文档内容以及根据词汇分布获知和某个特定词汇相关的章节部分。整个可视化界面是一个文档的概览图形，图形由两个同心的椭圆圈组成，将文档按行分析，外圈按照行的顺序列出了文档中的所有词，内圈列出了文档中所有不重复的词，出现频率越高的词，在内圈中显示的颜色会越明亮，内圈中词的位置就会越靠近其在文中出现最频繁的地方，如分布越均匀的词，会越靠近圆心。点击或检索一个词，图形中不仅会显示该词在外圈中的分布情况，同时原始文本视图中会高亮显示该词的分布，通过视图滚动条能查看其详细信息。

TextArc 很容易发现文档中词的使用规律，能清楚地知道词在文档的哪个部分出现较多。但是整个文档按行分析，没有逻辑结构的概念，无法体现出词出现在文档中不同位置的差异性。

此为试读，需要完整PDF请访问：www.ertongbook.com