计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(7):56−69 [doi: 10.15888/j.cnki.csa.007527]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于扩展 B_{cp} 指数的领域主题发展态势可视分析①

王　杨[1,2], 余敏楮[1,2], 单桂华[1], 田　东[1,2], 陆忠华[1]

[1](中国科学院 计算机网络信息中心, 北京 100190)
[2](中国科学院大学, 北京 100049)
通讯作者: 单桂华, E-mail: sgh@sccas.cn

摘　要: 通过对已发表论文的分析, 掌握研究领域的发展状况, 对研究人员具有重要意义. 面向此类需求, 提出一种基于扩展 B_{cp} 指数的领域主题发展态势可视分析方法. 首先, 从论文的标题、摘要以及作者提供的关键字中自动提取包含词组类型的关键词集合. 提取这些关键词之间的共现关系. 根据这些关键词使用 LDA 算法进行提取主题. 然后, 提出一种扩展 B_{cp} 指数来度量关键词的发展状态, 并据此对关键词和论文进行分类, 以确定发展状态类型. 基于此方法, 设计并实现了一个由需求驱动的领域主题发展态势可视分析工具 VISExplorer. 该系统可以展现领域主题分布和发展趋势、可以按主题推荐高质量文章、可以浏览不同主题中的高产出作者和高引用作者. 最后, 以可视化领域为例, 根据 1990 年至 2018 年在可视化领域顶级会议 IEEE VIS 上发表的论文, 对 VISExplorer 进行了实际案例应用, 并通过用户反馈证明了方法的实用性和有效性.

关键词: 文献计量学; 主题提取; 自然语言处理; 可视分析; 可视化

引用格式: 王杨,余敏楮,单桂华,田东,陆忠华.基于扩展 B_{cp} 指数的领域主题发展态势可视分析.计算机系统应用,2020,29(7):56–69. http://www.c-s-a.org.cn/1003-3254/7527.html

Visual Analysis for Development Situation of Research Topics Based on Extended B_{cp} Index

WANG Yang[1,2], YU Min-Zhu[1,2], SHAN Gui-Hua[1], TIAN Dong[1,2], LU Zhong-Hua[1]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: It is of great significance for researchers to master the development of research field through the analysis of published papers. In order to meet this requirement, a visual analysis method based on extended B_{cp} index was proposed. First of all, keywords containing phrases are automatically extracted from the title, abstract, and author provided keywords. Then co-occurrence relationship between these keywords was extracted. According to these keywords, LDA algorithm was used to extract topics. Then, an extended B_{cp} index was proposed to measure the development state of keywords. Based on this method, a visual analytic tool VISExplorer was designed and implemented. VISExplorer can show the distribution and development trend of domain topics, recommend high-quality papers, and browse top authors. Finally, taking the domain of visualization as an example, VISExplorer was conducted in real cases of publications on IEEE VIS Conference from 1990 to 2018, and the usefulness and effectiveness are proved by user's feedbacks.

Key words: scientometrics; topic extraction; natural language processing; visual analysis; visualization

① 基金项目: 中国科学院"十三五"信息化专项课题 (XXH13504)
Foundation item: CAS Special Fund for Informatization Construction in 13th Five-Year Plan (XXH13504)
收稿时间: 2019-12-23; 修改时间: 2020-01-20; 采用时间: 2020-02-11; csa 在线出版时间: 2020-07-03

通常, 某个学科领域的顶级会议和期刊上所发表的论文代表着世界在该领域内的最新研究成果. 该领域的研究人员都会对其中的前沿技术和高水平论文非常感兴趣. 因为这些会议和期刊所发表的论文代表着世界在该领域的最新研究成果. 他们时刻关注着该领域的研究主题及其研究趋势, 渴望了解其中高被引论文、热点主题和高度活跃的作者. 分析并掌握领域研究热点及前沿技术的发展态势, 对于科学家的研究工作、管理者的科技政策制定、甚至是研究生选题都具有重大的指导意义.

要研究领域主题的发展态势, 首要的问题就是如何从论文中提取领域中的主题. 主题可以用一组关键词来解释. 要提取领域中的主题, 本文需要首先获取关键词. 显然, 论文作者在其文章中提供的关键字是一个方便直接的来源. 然而, 有的论文并没有作者提供的关键字, 特别是在早期发表的论文中[1]. 还有相当一部分作者都认为有时作者提供的关键字并不能很好地表示论文的主题. 为了解决这些问题, 一种有效的方法是从论文的标题、摘要甚至全文中提取关键词. 然而, 单个单词的字关键词往往存在歧义. 例如, “network”一词既可能指社交网络也可能指神经网络. 因此, 也有必要提取包含词组形式的关键词而不是单个单词. 在获取领域关键词以后, 主题就可以通过用一组语义相关性高的关键词来定义, 即对关键词进行分类. 目前关键词提取及分类方法主要有两种, 一种是通过人工来筛选关键词并定义分类, 比如邀请领域专家来打分. 这种方式的优点在于精确度高、类别含义明确易懂, 缺点是普适性比较差, 每个会议、期刊或论文数据库都有自己的分类标准, 大多时候很难将一种来源的文献按另一种来源的分类方法一一对应. 当关键词数量庞大的时候, 人工方法的时间成本会变得巨大. 另一种方法是通过自然语言处理及聚类算法对关键词自动提取并聚类, 这种方法的优点在于普适性很强, 不论什么来源的论文, 都能通过一套算法自动实现提取及分类. 而且, 计算机算法在处理大量关键词的时候的具有人工无可比拟的优势. 其缺点在于提取到的关键词的质量跟算法的优劣有直接关系. 并且聚类结果是否有明确含义, 还需要人工进行验证.

对于领域主题, 现有的科学文献分析大都集中于使用传统的文献计量学方法, 如统计论文数量和被引情况, 建立被引用次数网络和合著网络等等. 本文需要通过更高阶的指数来揭示更深层次的现象和规律. 而高阶指数文献计量结合可视分析技术正是当前文献研究领域的热点研究方向之一.

本文的工作正是基于关键词提取、主题聚类、高阶指数计量和可视分析技术研究领域主题发展态势. 本文的主要贡献包括:

(1) 本文使用提取的词组而不是单词作为关键词. 这些词组是用自然语言处理的方法从标题和摘要中提取出来的. 基于这些关键词, 本文使用 LDA 和共现关系来研究领域论文中的主题分布.

(2) 将可视分析与文献计量学相结合, 分析领域主题的发展历史、现状和趋势. 本文提出一种扩展的 B_{cp} 指数用以描述发展状态, 并据此来判断一个主题或关键词发展状态. 同时, 本文将 B_{cp} 指数应用于判断一篇论文的被引用状态, 并将论文按引用状态分为“延迟承认”型、“长盛不衰”型以及“其他类型”. 在此基础上, 本文优化了经典的论文推荐方法. 本文还建立了一个作者的合作网络, 以便挖掘一个主题的研究社团.

(3) 本文开发了一个交互式可视化分析系统 VISExplorer, 支持科学文献的主题发展态势展示、趋势分析、社团发现和论文推荐.

1 相关工作

1.1 科学文献中的主题提取

主题提取技术已经被广泛地应用于文献分析. 典型的主题抽取技术包括共词分析和 LDA[2]的概率方法.

共词分析是根据关键字、标题、摘要乃至全文中的词的共现关系来提取主题的[3−7]. 与本文的工作最相关的研究有: Coulter 等[8]在软件工程领域的工作、Hoonlor 等[9]对计算机科学文献的普查工作、Liu 等[10]的基于人机交互的文献分析以及 Isenberg 等[1]对 IEEEVIS 论文数据的分析.

LDA 是 Blei 于 2003 年提出的, 是一种广泛应用于主题抽取和文本分类的概率模型. 许多工作[11−15]都致力于解释 LDA 提取的主题. Sievert 等[15]还开发了一个 LDA 模型的交互式可视化软件 LDAvis.

共词分析可以清楚地揭示关键词与主题之间的关系, 但这种方法主要依赖于人工对主题进行提取. 而使用 LDA 则更为方便, 也不需要太多的人工操作. 但是,

LDA 提取的主题可解释性不高. 本文中, 本文将这两种技术结合在一起. 本文用 LDA 从关键词中提取主题, 并用共词分析来显示主题和关键词之间的关系.

1.2 文献计量学文献分析方法

文献计量学中有关文献分析的经典方法包括被引用次数分析、共引分析、合著分析、影响力分析和评估等等. 本文将分析重点放在被引用次数分析和评估的基础上, 找出领域发展模式和重要的论文.

近年来, 在通过被引用次数寻找领域发展模式方面做了大量工作. 为了找到“延迟承认”模式的论文, Ke 等[16]系统地分析了自 20 世纪以来在自然科学和社会科学领域发表的 2200 多万篇论文的被引用次数. Van Raan 等[17,18]利用被引用次数分析研究了 1980-1994 年《Science》的被引用次数, 寻找物理、化学、工程和计算机科学领域的论文模式. Ke 等[16]提出了 B 指数来识别符合“睡美人”模式的论文. Du 等[19]扩展了 B 指数, 提出了一种 B_{cp} 指数, B_{cp} 指数能比 B 指数更准确地识别“延迟承认”类型的论文. 本文参考 Du 的思想, 提出一种扩展的 B_{cp} 指数来识别更多类型的论文.

1.3 科学文献的可视分析

Chuang 等[4]使用 Jigsaw[20]工具和 CiteVis 工具[21], 并基于 IEEE VIS 可视化论文的数据集 vispubdata[22], 构建了用于主题模型诊断的机器学习模型. Latif 等[23]开发了一个结合文本分析和可视化的交互式论文可视化系统, 以生成 IEEE VIS 论文的作者文字简介. Guo 等[24]使用迭代设计的可视化分析工具分析基于主题的意义构建框架和实验, 以确定主题设计的意义, 从而促进使用可视化生成研究想法. Federico 等[25]回顾了专利和论文的交互分析和可视化方法，并根据数据和任务两个方面对文献可视分析方法进行分类。

近年来, 与本文的工作类似的是 Isenberg 等[22]的工作. 基于作者提供的关键字, 他们展示了 1990~2015 年间发表在 IEEE 可视化会议系列 (现在称为 IEEE VIS) 上的论文的综合的多通道的分析结果. 他们对这些关键字进行了多次人工编码, 进而找到更高级别的关键字主题集合, 然后使用共词分析和策略图来研究主题的发展态势. 然而, 有将近 30% 论文没有作者提供的关键字, 他们只是简单地把这些论文从数据中剔除出去. 而且, 他们的工作依赖于大量人工编码工作, 这种分类只适合于研究 IEEE VIS 会议的论文, 对于其他刊源的数据集, 这种人工分类并不合适, 而且对于更大量的数据会耗费巨大的时间成本. 本文的方法是从标题和摘要中提取关键字, 将它们与作者提供的关键字相结合, 使用 LDA 代替人工工作提取主题, 运用文献计量学的方法对主题和论文模式进行识别. 相较而言, 本文的方法具有更高的效率和可扩展性.

2 需求分析

本文的用户群是处于研究生涯不同阶段的研究人员, 可以分为两类: 新手研究人员和经验丰富的研究人员.

新手研究人员是指那些刚开始自己研究生涯的研究人员. 他们正处于研究生涯的早期阶段, 对自己的研究领域了解不足. 他们迫切需要知道本领域包括哪些研究主题? 每个主题研究什么技术? 每个主题发展的历史和趋势是什么? 哪些文章是必读的关键文章? 等等. 这些信息可以帮助他们快速定位关键文章, 用最少的精力较深入地了解感兴趣的研究方向, 选择最合适的研究方向.

有经验的研究人员是指已经积累了某领域相当研究经验的研究人员. 他们正处于研究生涯的中期, 对自己领域内的各种研究方向有较深的理解. 这些研究人员基本都有一两个主要的研究主题, 他们经常需要这些主题的最新动态, 以寻找其中某些关键问题的解决方案. 他们需要知道这些主题是近几年的发展态势如何? 最活跃的作者有哪些? 有没有与自己的研究类似的重要论文发表? 这些信息有助于激发新的研究思路.

综上所述, 可以归纳出 4 个主要需求:

需求 1: 在宏观上展示主题分布. 用户可以在此基础上选择自己的感兴趣的研究主题, 进行深入了解和分析.

需求 2: 分析主题的发展趋势. 对于一个主题, 用户渴望了解该主题的研究热点以及相关重要论文. 因此, 需要一种有效合理的评价方法来评价该课题的发展态势.

需求 3: 显示每个主题中作者的合作关系. 一个领域的研究人员通常希望与该领域的其他同行进行交流, 尤其是对高被引或高产出的作者尤为关注. 此外, 研究社团可以帮助用户挖掘更多更精准的合作机会.

需求 4: 用户需要高效便捷地探索领域信息. 为了

满足上述要求, 需要一个交互式的可视化系统. 系统包含领域主题分布、趋势分析、作者合作关系和重要论文推荐等功能. 系统必须根据每次交互更新可视化内容, 以便用户能够实时获得聚焦主题的各维度信息.

3 数据处理

主题是本文分析的基本信息, 通常由作者提供的关键字表示. 然而, 并不是所有的论文都有这样的关键词, 特别是那些在 IEEE VIS 早期被接受的文献[22]. Isenberg 等发现, 2000 年以前 IEEE VIS 论文的关键词覆盖率不到 70%. 为了充分利用这 10 年的论文数据, 本文从论文的标题和摘要中提取关键词, 并在此基础上提取主题.

3.1 数据来源

本文收集了 1990~2018 年 IEEE-VIS 接收的 3067 篇完整论文. 这些论文数据来源于 vispubdata、IEEE VIS 官方网站、IEEE Xplore 和 Microsoft Academic. 每篇论文包括标题、作者、发表年份、会议、摘要、被引用次数等. 其中大部分论文包含了作者提供的关键字、IEEE 关键词、INSPEC 控制索引和 ISNPEC 非控制索引.

3.2 关键词提取

本文设计了一套关键词提取流程, 从标题和摘要中自动提取包含词组的关键字. 流程由 4 个主要模块组成.

M1: 预处理模块. 这一模块主要用于生成和清理用于提取关键词和主题的语料库. 将每一篇论文的标题和摘要合并为一个文档, 这样的 3067 个文档就构成了语料库, 并通过将所有单词统一为小写以及删除特殊字符等方法来清理语料库.

M2: 短语提取模块. 这一模块使用 NLTK 对语料库中的词性进行标记与分词. NLTK 是一个提供许多自然语言处理方法的 Python 库. 接下来, 基于 n-gram 模型生成 2-gram, 3-gram, ⋯, 6-gram 并提取名词词组. 这些名词词组, 与作者提供的关键字、IEEE 关键字、ISNPEC 的控制索引和非控制索引, 一起组成了关键词候选集. 鉴于在论文中的大多数核心关键词的长度都不超过 6 个单词, 本文将提取词组的最大长度设置为 6. 通过这种方法, 本文从 3067 篇论文中共提取出 6754 个核心关键词组.

M3: 共现矩阵生成模块. 这一模块计算关键词候选集中, 任意两个关键词的共同出现在一篇论文中的次数, 并将其存放到 6754×6754 大小的共现矩阵中.

M4: 关键词过滤. 这一模块根据过滤条件, 结合共现矩阵, 从关键词候选集中选择较重要的关键词, 将一些不重要的关键词过滤掉. 本文设置了 3 个过滤条件:

(1) 每个关键词都与一个以上的其他关键词有共现关系 (过滤掉孤点);

(2) 对于每个关键词, 包含它的论文数不小于 5 篇;

(3) 任意 2 个有共现关系的关键词的共现次数不小于 2 次.

经过过滤后的关键词就是本文所研究的领域关键词候选集. 通过少量的人工干预, 即可产生较高质量的关键词集合, 具体方案在第 5.2 节说明.

表 1 $\beta = 0.27$时的选词结果, 列出了每一个主题词频排名前 3 的代表词

主题	关键词
主题 1	medical image; biomedical image; image processing;…
主题 2	visual analysis; decision making; barchart;…
主题 3	flow visualization; vector field; computational fluid dynamic;…
主题 4	transfer function; volume rending; direct volume rendering;…
主题 5	information system; geographic system; web site;…
主题 6	surface reconstruction; graph layout; computational geoetry;…
主题 7	social network; network visualization; graph visualization;…
主题 8	time series; social medium; time series analysis;…
主题 9	visualization design; design study; visual encoding;…
主题 10	surface extraction; virtual reality; virtual environment;…
主题 11	weather forecasting; molecular visualization; view-dependent rendering;…
主题 12	diffusion tensor; diffusion tensor imaging; tensor imaging;…
主题 13	event sequence; geometric model; software package;…
主题 14	scalar field; principal component analysis; neural network;…
主题 15	text analysis; aspect ratio; tag cloud;…

3.3 主题提取

本文使用 LDA 模型从领域关键词集合中自动提取主题. LDA 是一种广泛应用于文本分类的基于概率的机器学习方法, 是一种典型的词袋模型. 它把一篇论文看作一个词袋, 词与词之间没有词序信息. 因此, 可以把一篇论文看作是由若干在论文中出现过的领域关键词所组成的词袋. 将这些论文词袋输入到 genism 库的 LDA 模型中, 并设置主题数量, 即可得到相应的主题.

用于投稿和评审论文的 Precision Conference System (PCS) 系统将关键词分成 14 大类, Isenberg 等人[1]在经过多名专家多次研讨后将关键词分成 16 类, 本文取平

均值, 将主题数量设为 15 个. 通过 LDA 模型得到了 15 个主题及其关键词分布, 并使用 Sievert[15]定义的显著性公式来选择每个主题的关键词:

$$r(w,k|\beta)=\beta\log(\varphi_{kw})+(1-\beta)\log(\frac{\varphi_{kw}}{p_w}),0\le\beta\le1 \quad (1)$$

其中, $r(w,k|\beta)$是关键词w和主题k的相关度. φ_{kw}是w属于k的概率. p_w是w在预料库中的边缘概率. β是平衡公式加号前后两部分的系数, 它是作为调节选词归属度优先还是词频优先的重要参数. $\beta=1$时, 选词标准就完全按照归属大小度选择. $\beta=0$时, 选词标准就变为完全按照词频大小选择. 表 1 是$\beta=0.27$时的选词结果, 列出了每一个主题词频排名前三的代表词.

4 基于文献计量学的主题发展态势分析

本节将阐述如何通过文献计量学方法来分析主题. 根据第 3 节的需求, 本文重点研究领域主题的发展态势. 主题发展态势是一个主题的研究历史和研究现状的表现, 主要反映在该主题相关的论文数量、论文质量、历年趋势、研究人员规模等指标上.

4.1 基于被引用趋势的主题/论文分类

本文把一篇论文的生命期定义为从论文发表时刻到当前时刻的这段时期. 一篇论文可以根据其生命期内的被引用次数分布情况来揭示其受关注程度. 同理, 一个主题的历年被引用次数可以通过将所有与该主题相关的论文的历年被引用次数相加来计算. 一个主题的历年研究热度变化反映在其生命期内的被引用次数分布情况. 论文/主题的历年被引用次数分布情况可以将分为 6 个子类型.

子类型 1: 引用集中在生命期的后期, 早期引用较少. 这说明, 论文发表/主题发展初期, 很少有人关注. 随着时间推移, 它的价值被慢慢发现, 并被大家广泛认可. 这意味着这篇文章或这类主题的研究内容可能是具有颠覆性或超前性的, 经过长期的沉寂, 在当前具有很强的研究价值. 图 1(a) 所示为子类型 1 的历年被引用曲线的示例形状.

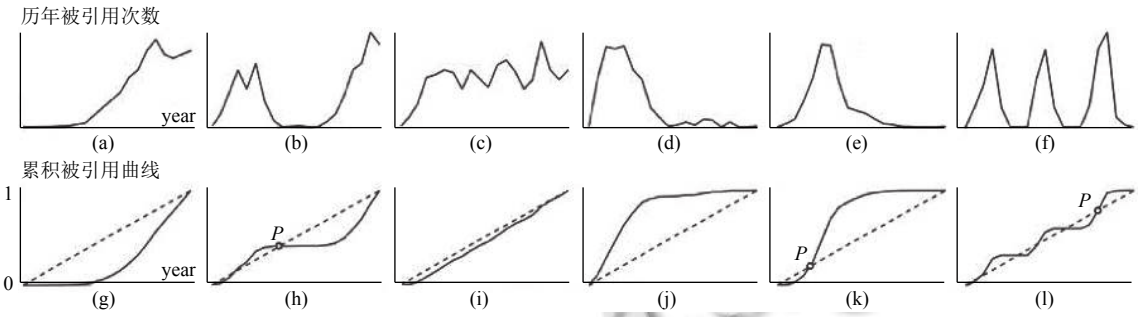


图 1 6 类被引用曲线形状和对应的累积被应用曲线形状

子类型 2: 引用集中在生命期的早期和晚期, 中期引用较少. 这意味着论文发表/主题发展之初就广受关注, 但随后关注度慢慢下降, 在沉默了一段时间后, 它又开始逐渐引起人们的注意. 这说明该论文/主题所涉及的研究内容在发表之初就显示出很高的研究价值, 但由于当时技术或知识上的不足, 相关研究遇到了瓶颈. 然而, 经过一段时期后, 由于知识的积累或技术的突破, 满足了继续推进研究的必要条件, 这些研究内容再次成为研究热点. 这种类型的论文/主题在当前也具有很大的研究价值. 图 1(b) 所示为子类型 2 的历年被引用曲线的示例形状.

子类型 3: 引用次数历年分布相对平均, 无大波动. 这说明论文/主题具有很强的生命力, 在其生命期内每年都能保持稳定的被引用率. 一般来说, 这些论文或主题所涉及的内容都是经典或基础的研究. 图 1(c) 所示为子类型 3 的历年被引用曲线的示例形状.

子类型 4: 引用集中在生命周期的早期, 后期的引用很少. 这表明论文/主题自发表以来受到了广泛的关注, 但随着时间的推移, 逐渐失去了人们的关注. 这意味着论文/主题中提到的研究内容现在已经过时、逐渐被遗忘, 或已达到成熟状态. 图 1(d) 所示为子类型 4 的历年被引用曲线的示例形状.

子类型 5: 引用集中在生命周期的中期, 早期和后期很少. 这意味着论文/主题在发表之初没有被注意到,

随着时间推移, 它的价值逐渐被发现和认识, 过了一段时间, 又失去了研究价值. 这意味着论文/课题中涉及的研究内容现在也已过时或研究已达到成熟. 图 1(e) 所示为子类型 5 的历年被引用曲线的示例形状.

子类型 6: 引用次数多次涨落, 波动较大. 在实际中, 只有总被引次数很少的论文/主题会出现这种情况. 那些重要的高被引文章或主题基本都不属于这种类型. 因此, 本文不予讨论. 图 1(f) 所示为子类型 6 的历年被引用曲线的示例形状.

这 6 个子类型还可以进一步合并为 3 大类:

第 I 类: 子类型 1 和子类型 2 的论文/主题总是包含最先进的技术或研究热点, 对研究人员最有价值. 这两种子类型的论文/主题的共同点是, 它们的被引用次数在生命期后期明显上升. 本文把这两个子类型合并成第 I 类.

第 II 类: 子类型 3 的论文/主题一般涉及基础知识或技术. 这对研究人员, 特别是新手研究人员也非常重要. 这类论文/主题的历年被引用情况相对稳定, 在生命期内没有显著的上升或下降趋势. 本文将子类型 3 归为第 II 类.

第 III 类: 子类型 4、子类型 5 和子类型 6 的论文/主题所包含的技术或知识通常是成熟的或过时的. 这类论文/主题的引用在生命期后期明显减少, 甚至消失. 本文将这 3 个子类型合并为第 III 类.

4.2 论文/主题类型识别

在第 4.1 节中, 我们根据论文/主题生命期内的被引用次数分布定义了 3 大类型和 6 个子类型. 但是, 如何通过数学方法自动判断一篇论文或一个主题属于哪一类? 在 Du 等[19]的研究中, 对子类型 1 的论文提出了一套基于累积被引用曲线的判别方法. 本文扩展了这一思想, 使之能满足判断所有类型.

对于任意时间段$[t_1,t_2], t_1<t_2$, 对于某一年$t\in[t_1,t_2]$, 一个论文/主题的历年累积被引用次数可以表示为:

$$f(t)=\sum_{i=t_1}^{t}C_i \tag{2}$$

在这个公式中, C_i表示论文/主题在第 i 年的被引次数, 由公式 (3) 可知, 论文/主题的历年累计被引次数单调递增. 当t_1是发表年份, t_2是当前年份时, $f(t_1)$是论文/主题发表年份的被引次数, 通常$f(t_1)=0$. $f(t_2)$是迄今为止该论文/主题的总被引用次数.

为了消除每篇论文总被引次数差距过大而产生的影响, 我们将式 (3) 除以$f(t_2)$进行标准化:

$$c(t)=\frac{f(t)}{f(t_2)} \tag{3}$$

式 (4) 就是本文接下来要重点研究的累积被引用曲线.

定义从$(t_1,c(t_1))$到$(t_2,c(t_2))$的直线为参考线, 用公式表述为:

$$l(t)=\frac{c(t_2)-c(t_1)}{t_2-t_1}(t-t_1)+c(t_1) \tag{4}$$

从定义可以看出, 与参考线相对应的论文/主题的历年被引用次数是恒定的. 也就是说, 如果一篇论文/主题每年有相同的被引用次数, 其累积被引用曲线与其参考线重合. 累积被引用曲线位于参考线上方的区域意味着该论文/主题的被引用次数总体趋势在此期间持续上升. 累积被引用曲线位于参考线以下的区域意味着该论文/主题的被引用次数总体趋势在此期间持续下降. 6 个子类型的累积被引用曲线的示例形状如图 1(g) 至图 1(l) 所示.

除去起点和终点, 累积被引用曲线与参考线的交点是论文/主题被引用次数从上升到下降或从下降到上升的转折点. 在本文中, 当提到“交点”时, 指的是除两条曲线的起点和终点之外的交点. 这些交点可分为两种类型:

A 型: 对于累积被引用曲线与参考线的交点$(t,c(t))$, t可能不是整数. 设t_i是整数年, $t\in[t_i,t_i+1]$. 如果 c, 则将交点$(t,c(t))$分类为 A 型. 例如图 1(h) 中的交点 P. A 型交点始终是论文/主题被引用次数的总体趋势即将由降到升的关键点, 即这类交点所对应的时间点往后一段时间内, 论文/主题被引用次数的总体趋势必然会上升.

B 型: 对于累积被引用曲线与参考线的交点$(t,c(t))$, 如果$c(t_i)<c(t)<c(t_i+1)$, 则将交点$(t,c(t))$分类为 B 型. 例如图 2(k) 中的交点 P. B 型的交点总是论文/主题被引用次数的总体趋势即将由从上升到下降的关键点, 即这类交点所对应的时间点往后一段时间内, 论文/主题被引用次数的总体趋势必然会下降.

基于上述这些定义, 就可以分析 I–III 类论文/主题的累积被引用曲线和参考线的特征. 为了便于表达, 将累积被引用曲线和参考线交点$P(t_p,c(t_p))$定义为靠近终点$(t_2,c(t_2))$的最后一个交点, 即最近一次发生趋势大变化的关键点. 如果累积被引用曲线和参考线没有交点,

则P就是起点$(t_1,c(t_1))$.

对于第 I 类: 其累积引用曲线 (带参考线) 如图 1(g)(h) 所示. 这一类的主要特点是: 累积被引用曲线在P与终点$(t_2,c(t_2))$之间的部分位于的参考线下方, 且这部分累积被引用曲线和参考线围成的区域面积较大. 如果有交点, 则P是类型为 A 的交点.

对于 II 型: 其累积被引用曲线 (带参考线) 如图 1(i) 所示. 这一类的主要特点是累积被引用曲线紧贴参考线或基本重合.

对于 III 型: 其累积被引用曲线 (带参考线) 如图 1(j)(k)(l) 所示. 不属于前两种类型的论文/主题都归为类型 III. 这一类的主要特点是: 累积被引用曲线在P和终点$(t_2,c(t_2))$之间的部分位于参考线上方. 如果有交点, 则 P 是类型为 B 的交点.

4.3 B_{cp} 指数

根据 Du 等[19]的研究, 为B_{cp}指数可定义为: 对于任何非零引用论文, $(c(t_2)-c(t_1))/(t_2-t_1)$是参考线$l(t)$的斜率. 对于任意$t\in[t_1,t_2]$, 计算$l(t)-c(t)$的值. 然后, 将这些值加在$t=t_1$和$t=t_2$之间, 得到$B_{cp}$指数.

指数可以用公式表示为:

$$B_{cp}=\sum_{t=t_1}^{t_2}\left(\frac{c(t_2)-c(t_1)}{t_2-t_1}(t-t_1)+c(t_1)-c(t)\right) \tag{5}$$

从式 (6) 可以看出, B_{cp}的值是累积被引用曲线位于参考线下的面积减去累积被引用曲线位于参考线上的面积. 因此, 若累积被引用曲线位于参考线下的面积大, 则$B_{cp}>0$, 反之, $B_{cp}<0$.

从累积被引用曲线上的点$(t,c(t))$到参考线的距离. $D(t)$可以定义为从该点到参考线的垂线段的长度. $D(t)$可通过以下公式计算:

$$D(t)=\frac{\left|\left(\frac{c(t_2)-c(t_1)}{t_2-t_1}(t-t_1)+c(t_1)-c(t)\right)\right|}{\sqrt{\left(\frac{c(t_2)-c(t_1)}{t_2-t_1}\right)^2+1}} \tag{6}$$

最大距离记为:

$$D(t_D)=Maximum(d(t),t\notin[t_1,t_2]) \tag{7}$$

注意到这时间不是被引用次数中变化最大的时间, 而是被引用次数累积到由量变产生质变的时间.

根据上述定义和公式, 我们可以通过B_{cp}指数来识别论文/主题的类型. 累积被引用曲线上最有趣的区域是最后一个交点 P 和终点$(t_2,c(t_2))$之间位于参考线下方的区域. 该区域表示近年来论文/主题的被引用次数呈上升趋势, 其所涉及的研究内容是热点.

对于 I 类论文/主题, 计算t_p和t_2之间的B_{cp}指数. 显然, $B_{cp}>0$, B_{cp}值越大, 面积越大, 说明上升期的持续时间或范围也越大. 为了区别于 II 型, 累积被引用曲线与参考线之间的最大距离$D(t_D)$不应太小. 所以本文设置了一个阈值来筛选$D(t_D)$, 此时$D(t_D)$大于阈值.

对于 II 类论文/主题, 其特点是累积被引用曲线紧贴参考线或几乎重合. 所以$D(t_D)$不应该太大. 此时$D(t_D)$小于阈值.

对于 III 论文/主题, 不符合前两种类型的论文/主题即为此类, 此时t_p和t_2之间的B_{cp}指数为负值, $D(t_D)$大于等于阈值.

表 2 中列出了这 3 类论文/主题的B_{cp}和$D(t_D)$的特征.

表 2 不同类型的B_{cp}特征

Type	involved technologies	features of citations	Intersections	$B_{cp}\in[t_p,t_2]$
I	state-of-the-art or hotspots	currently rising	0 or more, P is Type A	$B_{cp}>0, D(t_D)\geq threshold$
II	basic and fundamental	constant	0 or more	$D(t_D)<threshold$
III	mature or outdated	currently falling	0 or more, P is Type B	$B_{cp}<0, D(t_D)\geq threshold$

4.4 论文推荐

在众多论文中, 研究人员更关注那些高被引论文. 在高被引论文中, 研究人员更关注 I 类和 II 类论文. 这两类论文更具有重要的现实研究价值. 因此, 本文主要推荐第 I 类和第 II 类论文.

本文推荐第 I 类和第 II 类论文, 并按总被引用次数降序排列. 但是, 按照总被引次数降序排列存在不足: 被引次数较低的老文章可能会排在被引次数较低的新文章前. 如一篇发表了 20 年的文章被引 5 次, 一篇发表了 2 年的文章被引 5 次, 用户会更倾向于阅读后者. 因此, 设置了一个限制来优化推荐列表, 即每个推荐的论文必须满足以下两个条件之一:

条件 1. 这篇论文的总被引用次数足够高. 被高度引用的论文一直是研究人员最关心的论文. 高被引论文的定义根据实际需要而有所不同. 本文设置推荐论文的总被引次数不小于所有 I 类和 II 类论文的平均被引用次数.

条件 2. 这篇论文年均被引用次数足够多. 本文用年均被引用次数作为指标, 是因为对于新发表的论文(生命期≤5 年), 生命期很短, 总被引用次数不大, 将其与生命期长的论文相比没有意义. 因此, 为了消除生命期长短的影响, 尽可能推荐有价值的新发表论文, 本文设置推荐论文的年均被引用次数不小于所有 I 类和 II 类论文的平均年均被引用次数.

5 可视化设计

根据上述分析方法和思想, 本文设计实现了一个交互式可视化分析系统 VISExplorer. 如图 2 所示, 该系统由 6 个版块组成: 领域主题总览 (a)、关键词分布与分类 (b)、被引用趋势曲线 (c)、合著网络 (d) 和论文推荐 (e).

5.1 研究主题总览

主题和关键词是本文分析的基础. 用 LDA 模型提取的主题可以看作是高层次的主题, 而构成主题的关键词可以看作是低层次的主题. 主题的分布和趋势可以通过关键词的分布和趋势来反映. 因此, 本文使用主题和关键词作为切入点, 帮助用户找到他们想要的信息.

如图 2(a) 所示, 主题总览由 4 部分组成: a1 用于调整关联度β; a2 为主题选择区域; a3 显示所选主题的关键词分布, a4 为搜索框. 在 a2 中, 本文可以通过主题编号来选择某一主题, 该主题前 30 个最显著的关键词将显示在 a3 中, 并按显著性由大到小进行排序. 每次调整β, a3 将重新排序. 在 a4 中, 用户可以输入自己感兴趣的关键词进行模糊查询, 进而选择相关关键词进行下一步分析.

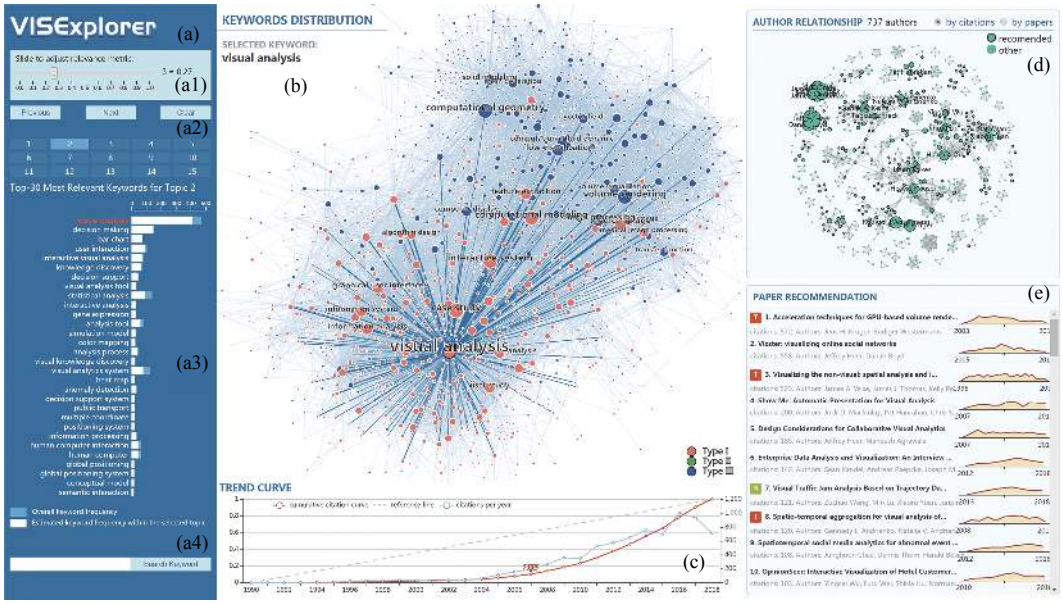


图 2 VISExplorer 系统界面

5.2 关键词分布和分类

为了使用户能够对整个 IEEE VIS 论文中所有主题的总体分布及关系一目了然. 我们需要清楚地展现两点: 关键词和主题之间的关系以及关键词之间的关系. 前者是展示 LDA 提取的主题结果. 后者是展示关键词内部的共现关系, 即共词分析.

基于上述考虑, 我们使用共词网络来表示关键词内部的关系, 如图 2(b) 所示. 每个节点代表一个关键词, 节点大小表示该关键词相关的论文数量. 两个节点之间的边表示这两个关键词有共现关系, 边的厚度与共现次数成正比. 根据本文提出的分类方法, 我们将所有关键词分类为 I、II、III 类, 并用不同的颜色来表示

不同的类型. 用户可以使用鼠标滚轮来放大和缩小图形, 也可以通过点击或圈选节点来选择他们感兴趣的关键词.

根据共现关系而形成的共词网络具有明显的聚类效果. 一个主题中具有相似语义或相似意义的关键字聚集在一起成为主题关键词群. 节点尺寸大的关键词表示了主题的主要研究内容, 并始终处于主题关键词群的中心附近. 不经常出现的关键词通常位于主题关键词群的边缘.

此外, 该共词网络可用于检验关键词提取效果. 本文基于 *n*-gram 模型提取关键词容易产生多余的关键词, 如 flow field visualization 关键词会产生 flow field 和 field visualization 关键词. 但在该共词网络中, flow field 和 field visualization 这类多余的关键词会紧紧围绕 flow field visualization 分布, 通过肉眼很容易发现. 因此, 通过该共词网络可以发现关键词提取过程中存在的问题, 辅助参数的设置, 以得到质量较好的关键词集合.

5.3 历年趋势

当用户选定关键字/主题以后, 将显示该关键字或主题的所有出版物每年的累积被引用曲线、参考线和历年被引用次数曲线. 这里我们使用双轴折线图来绘制趋势曲线, 如图 2(c) 所示. 在[0,1]范围内的左 Y 轴是累积被引用曲线和参考线的纵轴. 在 (0, +) 范围内的右 Y 轴是历年被引用次数曲线的纵轴. 这 3 条曲线共用一条表示时间跨度的 X 轴. 红色实线为累积被引用曲线, 灰色虚线为参考线, 蓝色实线为历年被引用次数曲线. 图中还使用针型图标来标记累积被引用曲线上到参考线距离最大的点.

5.4 作者合作网络

当用户选定关键词/主题以后, 本文采用力导向布局来展现其相关作者的合著网络, 如图 2(d) 所示.

图中每个节点表示选定主题/关键词的一个作者. 如果两位作者共同撰写了一篇该主题/关键词相关的论文, 则会在相应的节点之间连条边. 边宽与两位作者合著的论文数成正比. 本文采用两种不同的规则来映射节点的大小: 论文数量和被引用次数, 用户可以根据实际需求选择.

作者合著网络可以用来挖掘研究社区的分布. 由于同一篇论文的作者之间有相互关系, 这些作者的节点构成一个完全子图. 子图之间通过共同节点合并在一起, 形成更大的社区. 社区中节点越大, 代表的论文越多或被引用次数越多, 这些通常是社区中的核心专家. 如果某个节点作者的论文出现在论文推荐列表中, 则将该节点用黑色描边, 描边宽度与该作者被推荐的论文数量成正比.

5.5 论文推荐

当用户选定关键词/主题以后, 会在“论文推荐”版块中列出包含该关键词/主题的所有重要论文, 如图 2(e) 所示. 这些重要论文是根据 5.4 节中的方法对所有论文进行分类筛选后的结果. 图中同时也列出了论文的标题、被引用的次数、作者等信息, 并嵌入了每篇论文历年被引用次数曲线. 图中还使用含有字母的小图标来标记获奖论文或最近五年内发表的新论文. 标题前带有字母 T 的小图标表示本文获得了 IEEE VIS 大会的“Test of time”奖. 标题前面带有字母 B 的小图标表示该论文获得了当年的“Best paper”奖. 标题前带有字母 N 的小图标表示这篇论文是一篇最近五年内新发表的论文.

论文推荐列表使得用户可以轻松浏览相对重要和有价值的论文, 并根据曲线图观察论文历年被引用次数的变化.

6 案例分析

本文从 IEEE VIS 大会 1990~2018 年间收录的 3067 篇论文的标题和摘要中提取了 1799 个关键词和 15 个主题. 基于这些关键词和主题, 本节以真实案例为背景, 详细阐述如何通过 VISExplorer 来分析和展示可视化领域的主题分布、发展趋势、作者关系和重要论文.

6.1 关键词分布和分类

1799 个关键词及其共现关系如图 3 所示. 图中绿色的节点很少, 这说明第 II 类的关键词数量很少. 绝大部分关键词属于第 I 类和第 III 类. 从图中可以明显看出, 关键词分布有 3 个非常明显的聚类 (a), (b) 和 (c).

图 3 中 (a) 区域具有代表性的关键技术是尺寸较大的节点, 包括: visual analysis、case study、user study、information analysis 等, 这些关键技术基本上都属于信息可视化和可视分析范畴.

图 3 中 (c) 区域具有代表性的关键技术包括: volume rendering、computational geometry、flow visualization、vector field、medical image processing、

computational dynamic 等, 这些关键技术基本上都属于科学可视化范畴.

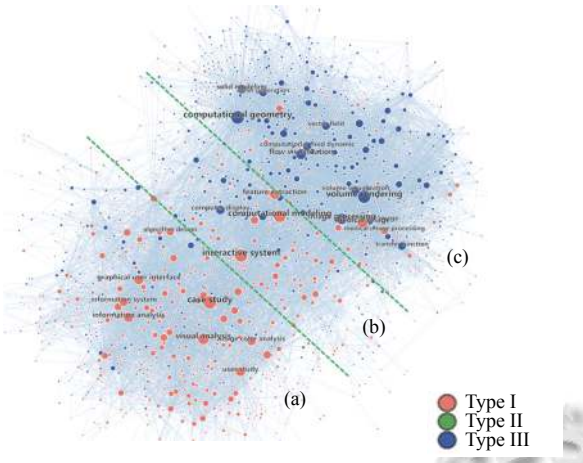


图 3 显示 IEEE 可视化会议中关键字的分类和分布, 按 3 类分类进行分析, 并通过共词分析进行链接

图 3 中 (b) 区域具有代表性的关键技术包括: interactive system、computational modeling、feature extraction、computer display 等, 这些关键技术基本上都属于可视化共性技术.

从图 3 中还可以看出, (c) 区域中的节点几乎都属于第 III 类, 这意味着近年来对传统科学可视化技术 (如体绘制、矢量场和特征提取) 的引用在下降. 这表明科学可视化的大部分技术的研究已经逐渐成熟或者遇到瓶颈. 同时, 医学图像处理 (medical image processing) 的节点为 I 型, 这意味着医学图像处理在当前仍然保持着良好的研究热度. 区域 (a) 中的节点大多为第 I 类, 这说明目前在信息可视化和可视分析领域的研究热度普遍很高. 区域 (b) 中的第 I 类和第 III 类节点数量差别不大, 所以对于可视化共性技术而言, 其研究热度相对平稳. 交互系统 (interactive system)、特征提取 (feature extraction) 和计算建模 (computational modeling) 是当前可视化共性技术的研究热点.

6.2 关键技术: Volume rendering (体绘制)

本文首先选择 volume rendering (体绘制) 作为第一个案例进行深入分析. 图 4 显示了体绘制技术的趋势曲线. 通过累积被引用曲线, 可以看出累积被引用曲线与参考线之间的最大距离发生在 2003 年. 这表明, 2003 年以后, 体绘制论文的引用量发生了质的飞跃. 2012 年前后, 累积被引用曲线与参考线产生交点, 这表明自此以后, 人们对体绘制技术的研究兴趣逐渐减弱. 历年被引用次数曲线证实了这一趋势.

图 4 volume rendering (体绘制) 相关论文的累计被引用曲线、参考线和历年被引用次数曲线

从图 4 中, 可以看出体绘制技术的发展经历了 3 个阶段.

第 1 阶段为 1990~2003 年. 在这一阶段, 体绘制技术经历了技术积累期. 在这一阶段, 其相关论文的被引用次数逐年增加.

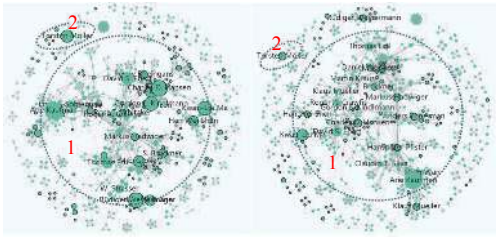
第 2 阶段为 2004~2012 年. 在这一阶段, 体绘制技术经历了一个繁荣时期. 其相关论文的被引用次数量在这一阶段初期迅速上升, 并在之后继续保持高被引用状态.

第 3 阶段从 2013 年开始至今. 在这一阶段, 大多数的体绘制技术研究日趋成熟或者遇到瓶颈, 有些可能已经过时. 其相关论文的被引用次数逐渐下降.

图 5(a) 和图 5(b) 显示了所有发表过体绘制相关论文的作者的合著网络. 图 5(a) 中的节点大小表示被引用次数, 图 5(b) 中的节点大小表示论文数. 可以看出, 图 5(a) 和图 5(b) 具有相同的网络结构. Arie E. Kaufman、David S. Ebert、Charles D. Hansen、Tomas Ertl、Han Wei Shen 和 Kwan Liu Ma 等构成了与体绘制相关的主要研究社区, 如图 5(a) 和 (b) 中的区域 1. 他们之间的合作程度、相关的论文数和被引用次数都很高. 其他较小的社区, 如图 5(a) 和图 5(b) 中的区域 2 所示, 如 Torsten Móller 社区, 也有大量的论文和被引用次数.

图 6 显示了根据第 4.4 节中阐述的规则推荐的体绘制相关的前 20 篇重要论文. 其中, 第一篇论文“Acceleration

techniques for GPU-based volume rendering”于 2018 年获得 Scivis“Test of time”奖. 从列出的 20 篇论文中, 可以看到, 这些论文都至少是 10 年前出版的.



(a) 节点大小表示被引用次数 (b) 节点大小表示论文数

图 5 体绘制相关作者的合著网络

6.3 关键技术: Visual analysis (可视分析)

本文选择 visual analysis(可视分析) 作为第二个案例进行深入分析. 图 7 显示了可视分析技术的趋势曲线. 通过图 7, 可以看到从累积被引用曲线到参考线的最大距离发生在 2007 年. 这表明, 2008 年以后, 体绘制论文的引用量发生了质的飞跃, 比相同的体绘制质变时间晚了 5 年. 而在整个可视分析的生命期中, 累积被引用曲线与参考线之间没有交点, 说明可视分析技术的被关注度一直在增长. 历年被引用次数曲线也证实了这一趋势.

从图 7 中可以看出, 可视分析技术的发展经历了两个阶段.

图 6 Volume rendering 相关的前 20 推荐文章

图 7 Visual analysis (可视分析) 相关论文的累积被引用曲线、参考线和历年被引用次数曲线

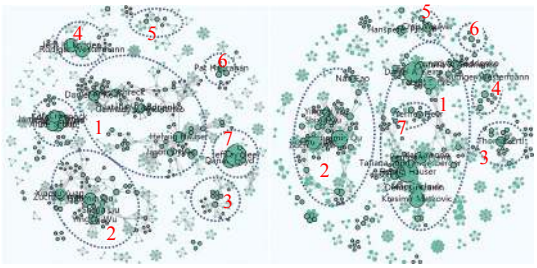
第 1 阶段为 1990~2007 年. 在这一阶段, 可视分析经历了长期的技术积累. 将近 15 年, 可视分析技术每年的被引用次数都不高.

第 2 阶段从 2008 年开始至今. 在这一阶段, 可视分析技术经历了它的繁荣时期. 在这一时期内, 相关论文的被引用次数逐年迅速上升. 越来越多的研究人员发现并认识到可视分析的重要性, 相关技术发展迅速, 受到越来越多的关注和应用.

图 8(a) 和图 8(b) 展示了发表可视分析相关论文的所有作者的合著网络. 图 8(a) 中的节点大小表示被引用数量, 图 8(b) 中的节点大小表示论文数量.

从图 8(a) 和图 8(b) 中, 可以看到可视分析中有两个相对较大的社区. Helwig Hauser、Kresimir Matkovic、Daniel A. Keim、Tobias Schreck 等构成了最大的社区, 如图 8(a)(b) 区域 1 所示. Huamin Qu、Xiaoru Yuan、Shixia Liu 和 Yingcai Wu 构成了第二大社区, 如图 8(a)(b)

区域 2 所示. 两个社区内的作者高度合作. 这两个社区都有大量的论文和被引用次数.



(a) 节点大小表示被引用次数 (b) 节点大小表示论文数

图 8 Visual analysis 相关作者的合著网络

一些小社区也有大量的论文和引用, 例如 8(a)(b) 区域 4. 有些社区发表了许多论文, 但引用率不高, 如 8(a)(b) 区域 3 和区域 5. 而 8(a)(b) 区域 6 和区域 7 则获得了较高的被引用次数, 却没有发表很多的论文. 图 9 显示了根据第 4.4 节中阐述的规则推荐的可视分析相关的前 20 篇重要论文. 除第 1 篇论文外, 第 3 篇论文“Visualizing the non visual spatial analysis and interaction with information from text documents”, 曾在 2016 年获得了 Inforvis 的“Test of time”奖. 第 8 篇论文“Spatio-temporal Aggregation for Visual Analysis of Movements”, 获得了 2018 年“Test of time”奖. 值得注意的是, 在这 20 篇论文中有 13 篇是在最近 10 年内 (2008 年之后) 发表的, 其中 3 篇是在最近 5 年内发表的, 这意味着可视分析技术的更新速度远远快于体绘制技术.

图 9 visual analysis 相关的前 20 推荐文章

6.4 用户反馈

为了评估 VISExploer 的实用性和有效性, 本文邀请了可视化领域的研究人员对本文的系统进行实用测试. 这些人中包括学生、教师、教授. 每个人都在使用后写了对系统的反馈, 并提出了大量很有价值的建议. 本节将列出其中两条反馈.

反馈 1: “通过选择主题和关键词, 我可以了解关键词之间的关系、发展状况和值得阅读的论文列表. 与现有的通用搜索引擎或文献检索库相比, 系统推荐的论文列表更具代表性. 作为对可视化领域尚了解不深的新手, 我可以通过阅读经典论文来了解可视化. 推荐论文列表中的论文都是最具里程碑意义的论文, 可以防止我盲目地在文档库中搜索, 从而节省大量的时间和精力. 此外, 我建议增加对新发表的综述型论文的推荐, 这样可以帮助新手快速了解可视化技术.”

反馈 2: “主题趋势分析和作者网络与实际需求密切相关. 论文推荐也很有意义. 这个系统不仅推荐了具有里程碑意义的老文章, 而且推荐出了优秀的新文章. 很感激. 作者网络可以快速定位领域专家并观察他们之间的合作情况. 我的建议是, 这个系统可以增强关键字搜索的功能. 允许用户根据自己的兴趣或实际需要自由搜索各种关键字组合. 此外, 如果系统能够支持更多的论文数据源, 那就更好了.”

7 总结与展望

本文提出了满足领域主题发展态势分析相关实际问题的解决方案, 并在此基础上开发了一个交互式可视化分析系统 VISExplorer, 并利用该系统, 对 IEEE

VIS 大会 1990~2018 年收录的 3067 篇论文的主题发展态势进行了研究. 本文还邀请了不同类型的研究人员来评估 VISExplorer 系统. 分析结果和用户反馈证明了该系统的有效性和实用性.

本文的工作仍存在一些局限性. 首先, 由于一个领域、一个主题或一篇论文都可能涉及多种技术. 本文以关键词提取算法来提取关键词, 在关键词质量上是不够的. 因此, 在未来工作中, 我们需要设计一个自动关键词检测系统, 将关键词提取算法辅以可视分析技术来提炼高质量的关键词. 第二, 本文只从标题和摘要中提取关键词, 这可能不能完全反映论文所涉及的所有关键技术, 因为并不是论文的所有关键技术都会出现在标题和摘要中. 因此, 今后我们将尝试以论文全文作为语料库进行关键字提取. 第三, 我们需要研究更多论文类型识别方法, 用以识别特别类型的论文, 如评论、综述等等, 这将有助于用户获取更精准的建议.

参考文献

1 Isenberg P, Isenberg T, Sedlmair M, *et al*. Visualization as seen through its research paper keywords. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 771–780. [doi: 10.1109/TVCG.2016.2598827]

2 Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993–1022.

3 Callon M, Courtial JP, Laville F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. Scientometrics, 1991, 22(1): 155–205. [doi: 10.1007/BF02019280]

4 Chuang J, Gupta S, Manning CD, *et al*. Topic model diagnostics: Assessing domain relevance via topical alignment. Proceedings of the 30th International Conference on International Conference on Machine Learning. Atlanta, GA, USA. 2013.612–620.

5 He Q. Knowledge discovery through co-word analysis. Library Trends, 1999, 48(1): 133–59.

6 Law J, Bauin S, Courtial JP, *et al*. Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification. Scientometrics, 1988, 14(3–4): 251–264.

7 Kostoff RN. Co-word analysis. Bozeman B, Melkers J. Evaluating R&D Impacts: Methods and Practice. Boston: Springer, 1993.63–78.

8 Coulter N, Monarch I, Konda S. Software engineering as seen through its research literature: A study in co-word analysis. Journal of the American Society for Information Science, 1998, 49(13): 1206–1223. [doi: 10.1002/(SICI)1097-4571(1998)49:13<1206::AID-ASI7>3.0.CO;2-F]

9 Hoonlor A, Szymanski BK, Zaki MJ. Trends in computer science research. Communications of the ACM, 2013, 56(10): 74–83. [doi: 10.1145/2500892]

10 Liu Y, Goncalves J, Ferreira D, *et al*. CHI 1994–2013: Mapping two decades of intellectual progress through co-word analysis. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Toronto, ON, Canada. 2014. 3553–3562.

11 Almanza M, Chierichetti F, Panconesi A, *et al*. A reduction for efficient LDA topic reconstruction. Advances in Neural Information Processing Systems. Montreal, QB, Canada. 2018. 7869–7879.

12 Newman D, Noh Y, Talley E, *et al*. Evaluating topic models for digital libraries. Proceedings of the 10th Annual Joint Conference on Digital Libraries. New York, NY, USA. 2010. 215–224.

13 Taddy M. On estimation and selection for topic models. Proceedings of the 15th International Conference on Artificial Intelligence and Statistics. La Palma, Spain. 2012. 1184–1193.

14 Chuang J, Manning CD, Heer J. Termite: Visualization techniques for assessing textual topic models. Proceedings of the International Working Conference on Advanced Visual Interfaces. New York, NY, USA. 2012. 74–77.

15 Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Baltimore, MD, USA. 2014. 63–70.

16 Ke Q, Ferrara E, Radicchi F, *et al*. Defining and identifying sleeping beauties in science. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(24): 7426–7431. [doi: 10.1073/pnas.1424329112]

17 Van Raan AFJ. Dormitory of physical and engineering sciences: Sleeping beauties may be sleeping innovations. PLoS One, 2015, 10(10): e0139786. [doi: 10.1371/journal.pone.0139786]

18 Van Raan AFJ. Sleeping beauties in science. Scientometrics, 2004, 59(3): 467–472. [doi: 10.1023/B:SCIE.0000018543.82441.f1]

19 Du J, Wu YS. A parameter-free index for identifying under-cited sleeping beauties in science. Scientometrics, 2018, 116(2): 959–971. [doi: 10.1007/s11192-018-2780-0]

20 Görg C, Liu ZC, Kihm J, *et al*. Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(10): 1646–1663. [doi: 10.1109/TVCG.2012.324]

21 Stasko J, Choo J, Han Y, *et al*. Citevis: Exploring conference paper citation data visually. Posters of IEEE InfoVis. Atlanta, GA, USA. 2013. 2.

22 Isenberg P, Heimerl F, Koch S, *et al*. Vispubdata. org: A metadata collection about IEEE visualization (VIS) publications. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(9): 2199–2206.

23 Latif S, Liu D, Beck F. Exploring interactive linking between text and visualization. EuroVis (Short Papers). Short Papers, Brno, Czech Republic. 2018. 91–94.

24 Guo H, Laidlaw DH. Topic-based exploration and embedded visualizations for research idea generation. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(3): 1592–1607. [doi: 10.1109/TVCG.2018.2873011]

25 Federico P, Heimerl F, Koch S, *et al*. A survey on visual approaches for analyzing scientific literature and patents. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(9): 2179–2198. [doi: 10.1109/TVCG.2016.2610422]