交互式文献可视化系统设计与实现[i]

李胜嘉

南京工程学院 211167

摘要：【目的】从文献分析的角度出发，设计并实现了一个将文献机构作者数据分析可视化的一个交互式文献可视化系统原型，在一定程度上弥补了现有文献可视化工具的缺陷。**【方法】**首先介绍文献可视化相关领域的国内外研究现状，然后对实现文献可视化系统的整体思路与功能架构进行设计，最后通过可视化领域的关键词，实现了文献的可视化与分析。**【结果】**本系统可以满足科研工作者进行初级文献可视化分析并了解某一学科领域的研究热点的需求，相较于传统的可视化工具有一定的体验优化。

关键词：可视化；信息系统；文献分析

Design and implementation of interactive paper visualization system

Abstract:

Objective: From the perspective of literature analysis, an interactive literature visualization system prototype is designed and implemented to visualize the data analysis of authors of literature institutions, which makes up for the defects of existing literature visualization tools to some extent.**Method:**This paper firstly introduces the research status of literature visualization related fields at home and abroad, then designs the overall thinking and functional framework of literature visualization system, and finally realizes literature visualization and analysis through keywords in the visualization field.**Conclusion:**This system can meet the needs of researchers to conduct visual analysis of primary literature and understand the research hotspots in a certain discipline. Compared with the traditional visual chemical industry, it has certain experience optimization.

Key Words: Visualization;Information system;Literature analysis

1.引言

随着科技的进步与高等教育的快速发展，一方面科研工作者数量急剧上升，另一方面，科学研究领域的重要成果——科研文献的数量也在急剧增长，如何有效地对科研文献进行分析，快速获取目标信息，成为文献分析领域的研究热点。“一图胜千言”，文献可视化可以满足科研工作者日益增长的个性化、多样化的文献研究需求，揭示文献背后隐藏的网络关联及规律，提高文献研究效率[1]。传统上，科研工作者对相关领域文献进行可视化操作往往要手动从期刊全文数据库获取相关文献题录信息并导入主流文献可视化工具进行分析，但是，科

[i] 本文为南京工程学院挑战杯项目（编号：TP20190006）资助研究成果。

研工作者想要从期刊数据库海量文献中检索到所研究内容并不容易，且由于计算机软件缺乏概念理解能力不能进行语义关联和推理。同时，使用传统文献可视化工具要求科研工作者具有一定的理论基础及计算机操作能力，这无疑加大了科研工作者进行文献研究的难度，降低了科研工作的效率。因此，在文献分析的过程中，怎样快速低成本的获取大量文献，并全面而准确的发现某一学科领域的研究热点和趋势[2]？对于特定的学科领域，怎样充分的利用文献信息并全面的展示？对于缺乏相关理论基础的科研人员，怎样降低文献研究的认知负担？以上问题的解决，有助于科研工作者了解研究热点，降低文献研究所需门槛，对文献研究乃至科研进步与发展都具有重要的现实意义。与此同时，笔者通过文献调研发现，图书情报与数字图书馆、教育理论与教育管理、计算机软件及计算机应用等领域的学者均层对此问题进行研究，并且开发出了一系列具有实用价值的文献可视化工具，在这之中较为知名的有陈超美博士等开发的科研文献分析系统 CiteSpace 以及由 Vladimir.Batagelj 和 Andrej.Mrvar 应用 Delphi 语言于 1996 年共同开发的一款用于分析大型复杂网络的软件 Pajek[3]。通过对以上工具对比研究，笔者发现此类工具存在着诸多问题，主要表现为：功能不够完备，只具有展示作者信息或者领域研究信息等某一方面的功能；信息利用不充分，在对机构或作者进行分析时没有考虑到所有属性；视图过于混乱，对于用户来说认知负担大，且体验较差。

基于此，笔者围绕上述问题展开研究，基于期刊论文数据数据库，结合地理信息可视化、网络数据可视化、文本内容可视化等信息可视化技术，设计并实现了将文献机构作者数据分析可视化的一个交互式文献可视化系统原型，可以对机构分布、发文量趋势、被引量趋势、合著网络、关键词文本等数据进行分析。本文首先介绍与文献可视化相关领域的国内外研究现状，然后对实现文献可视化系统的整体思路与功能架构进行详细解释，最后，对本系统所存在的缺陷以及未来改进的方向进行展望。

2.相关研究

2.1 文献可视化在国外的研究

文献可视化系统是伴随着科研文献数量快速增长，为了便于科研工作者进行初步的文献可视化分析并了解相关学科领域研究热点及趋势且结合多种信息可视化技术而产生的信息分析系统。文献可视化相关研究在国际范围内成为热点研究始于上世纪 90 年代中期，被公众所知为 IEEE 研讨会。始于 1995 年，每年 10 月召开于美国的的 IEEE 信息可视化专题研讨会结束后均会出版一系列会议论文集，论文集研究成果在业内产生了广泛而深远的影响。国外的文献可视化研究已经具备了初步的研究成果，在理论领域较为注重文献可视化模型方法，而在应用领域，不但涌现出一批初步的、低层次的原型系统，而且存在部分已经投入

使用的原型系统。

笔者以主题为“文献可视化”在 EBSCO 外文数据库进行检索，检索所使用的数据库有 MEDLINE,Business Source Ultimate,Academic Search,Eric,Library,Information Science&Technology Abstracts。检索范围为 1983-2018，检索式为“Title= (visualization) AND Abstract= (paper research)”，检索结果为 164 篇。EBSCO 数据库的检索结果表明，最早关于文献可视化的外文文献是 1983 年 Braden，RA 和 Walker，AD 所撰写的“Seeing ourselves: visualization in a social context. Readings from the Annual Conference of the International Visual Literacy Association (14th)[4]”，此篇为一书籍。笔者通过研读以上文献发现，国外在信息可视化的研究方向主要集中于医学可视化、可视化工具在文献关系揭示中的研究、可视化检索模型研究、个人信息搜索并可视化显示等，说明可视化为文献检索提供可视的直观的效果，同时文献检索也促进了可视化技术的不断发展。

2.2 文献可视化在国内的研究

我国最初几年关于文献可视化的研究中大量的文献主要是对信息检索可视化概念的介绍，以及地理、空间、数据库、文献和多媒体等方面的可视化技术的介绍。现在已经有越来越多的学者关注面向网络及大规模文献检索可视化技术、各个具体领域的可视化、各种算法的改进等具体问题。针对可视化技术在文献研究领域中的应用主要有:胡志刚、侯海燕撰文对科学技术期刊群中的 17 种期刊进行了聚类和社会网络分析，可视化地显示了各个期刊之间的亲疏关系和关联特点，其可视化对象是期刊及其期刊之间的关系，其研究对象是期刊论文[5]。鲍杨、朱庆华在论文中以 CSSCI 数据库收录的全部情报学领域的论文 (1998—2007 年) 为数据源，运用社会网络分析方法，建立了较为完整的情报学研究领域引文网络、共引网络及作者合著网络。同时他们还选取其中的重要节点，用 Pajek 进行了可视化，体现了近 10 年来我国情报学研究领域的主要作者和论文[6]。张学福教授在信息检索可视化领域的研究成果颇多，他不仅介绍了信息检索可视化的基本问题，即信息检索模型、信息内容描述、可视化映射技术、可视化显示技术、全局映射与局部映射、实时可视化和人工参与的可视化等。而且从功能特点等角度介绍了几种代表性的可视化开发工具:Open GL、Open Inventor、IDL 和 VTK 等，以便用户根据其特点选择相应的开发工具来开发可视化信息检索的应用软件[7]。张学福教授带领的研究生有三篇相关的硕士论文:《基于引文的信息检索可视化系统研究》一文基于信息检索可视化技术及引文理论，研究并构建了具有个性化特色的集检索、可视化及统计分析功能于一体的基于引文的信息检索可视化系统[8];《基于摘要信息的中文信息检索可视化系统研究与实现》分析比较国内外典型的信息检索可视化系统，并将基于词共现的概

念空间方法与信息检索可视化技术相结合实时生成概念空间图，实现了检索过程和检索结果的可视化，设计并实现了集成信息检索、情报分析和服务功能的基于摘要信息的中文信息检索可视化系统，并对系统进行测试与评估[9]。

3.系统思路与构建

3.1 系统构建

综合考虑现有及类似系统的功能与需求，本文实现的文献可视化系统大体分为五个模块：检索模块、导航模块、指数分析模块、计量可视化模块、资源分布模块，系统功能框架见图 1。

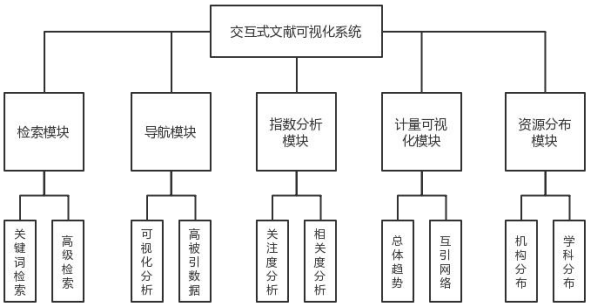


图 1 系统功能框架

3.1.1 检索模块

交互式文献可视化系统的检索模块实现了基于文献的关键词检索，该模块根据用户输入的关键词检索并返回符合条件的文献信息列表。检索模块满足用户简单快捷的获取感兴趣的文献关键词，有效地克服期刊文献数据库数量庞大、处理复杂而带来的问题。检索模块的实现是基于 TextRank 的关键词提取算法实现的，它分为关键词提取阶段和关键词检索阶段。关键词提取阶段是在信息管理模块中的信息添加操作中，使用 TextRank 算法提取文献信息的关键词，将提取的结果存储在 mysql 数据库中。

3.1.2 导航模块

用户在使用本系统了解相关学科领域的研究热点时，往往希望了解该学科领域的所有研究热点。导航模块的作用就是帮助用户以浏览的方式查找某一领域内的所有研究问题和研究方法。用户可以在导航栏“高被引数据”中查看以作者、期刊、院校、医院、文献、学科为分类的所有高被引数据。同时，高被引数据提供了被引率、H 指数等多种指标来衡量文献资源的价值，降低了使用者查找相关学科领域影响力较高的顶尖文献的成本。

3.1.3 指数分析模块

用户在进行关键词检索后，系统会依据该关键词对数据进行预处理，经过数据库存储等操作后将会呈现文献可视化页面。系统的指数分析模块包括学术、用户、媒体关注度及学术传播度等内容。在每个生成的可视化界面中，用户均可以在折线图、柱状图、堆叠图中自由选择，以满足不同需求的文献可视化需要。同时，系统也提供了生成图表的原始数据，以供科研工作者对此类数据进行进一步的分析。在完成文献关键词的可视化后，用户可以直接将可视化结果以图片的形式保存，并可以不加处理直接用于报告、论文的写作。

3.1.4 计量可视化模块

该模块主要是将传统的文献计量的结果用图表的方式表示。统计的对象包括每年发表文献数量、作者发表的文献数量、机构发表的文献数量、期刊发表的文献数量、引文数量和被引用期刊数量等，这些统计结果都以柱状图和折线图两种形式展现。该模块位于第二个选项卡，分为两个部分，第一部分是文献集合的一些基本统计信息，包括:总文献量、总作者量、总作者单位量、总期刊量、总关键词量、总引文量、总被引作者数量、总被引期刊数量等信息。除此之外，包括一些延伸的信息:每个作者平均发文的数量、每个机构发文的数量、每个期刊发文的数量、每篇文献的平均引文数量、平均引用的期刊数量等信息。第二部分是一些统计结构的可视化，是使用了开源图表工具 JFreeChart 实现的，统计图表可以输出为图片格式的文件，方便用户保存和使用。

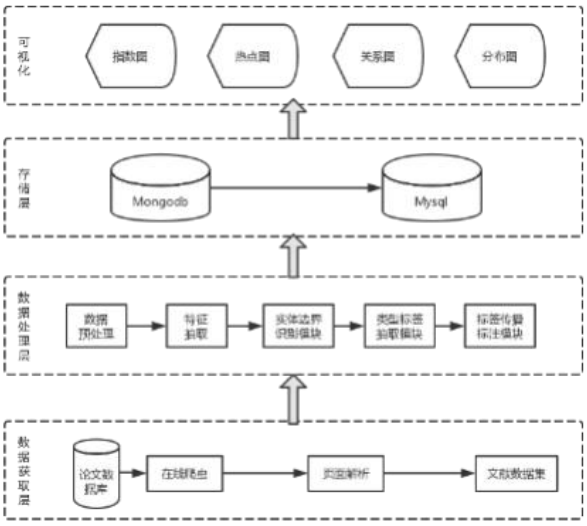
3.1.5 资源分布模块

该模块把通过关键词获取的文献信息以资源类型、学科、来源、基金、作者、机构分布为分类进行可视化展示，同指数分析模块近似，该模块也提供了不同图表呈现、原始数据保存等功能。

3.2 系统架构

如图 2 所示，文献可视化系统分为 4 层逻辑结构，自底向上分别是:数据获取层、数据处理层、存储层和可视化层。各层的功能依次递进，紧密相扣:最底层为数据获取层，它的功能是数据源的获取和存储，主要包括在线爬虫及其管理模块、页面解析模块和本地文献存储模块；数据处理层,主要为上面两层提供核心处理算法，包括有知识实体边界识别、类型标签抽取方法及基于多标签加权标签传播的类型标注方法等关键技术的实现；中间的存储层主要是把处理后的数据进行数据库存储并建立索引，将相关文献初始为 xml 格式，将其全部存入非关系型数据库 mongodb 中。利用 beautifulsoup 对初始文献内容进行解析，以获取研究需要的信息，将解析结果存入关系数据库 mysql。然后对数据进行知识实体关系图建模，并转换成 JSON 格式数据实现可视化；最上面的可视化层主要是与用户进行可视化交互，功

能是根据用户的输入反馈出不同的可视化关系图，包括有指数图、关系图、热点图等。以下



对部分重要模块进行介绍。

图 2 系统架构图

3.2.1 在线爬虫模块

管理员可以通过后台指定爬取页面的地址和范围，在线爬虫模块在后台自动化地对文献数据进行爬取并存储在本地，从而实现定向的文献爬取及分析。这样可以简便地把实体类型抽取扩展到其他专业领域或者其他论文数据库，为上面三层提供了充足的数据来源。

3.2.2 类型标签抽取及类型标注模块

知识实体的类型标签抽取模块主要是对识别到的知识实体进行类型标签抽取，得到类型标签集合和部分标注数据。然后，通过基于多标签加权的标签传播算法对未标知识实体实现进一步的标注，得到的类型标注数据传递给存储层进行本地存储，并建立知识实体及其类型关系索引库，提高检索效率。

3.2.3 模型构建及可视化模块

为了更好地将挖掘到的知识脉络实现可视化，我们需要对知识实体及其类型数据进行图模型的构建。根据用户输入的关键词对索引库进行检索，构建出不同的知识实体关系图模型，包括有基于同一类型的实体层次关系树模型 (层次图) 、基于不同类型分组的知识关系图模型 (关系图) 和基于时序的知识热点跟踪图模型 (热点图) 。然后，把得到的关系图模型转换成 JSON 格式的数据，传递到应用层利用 Echarts 进行 Web 可视化实现。

4.系统实现

4.1 爬虫实现

4.1.1 爬虫流程

系统主要利用 Httpcrawl 的 Bio.Httpcrawl 模块以编程方式访问 CNKI，用 java 脚本实现对相关数据库的搜索以及数据下载，批量抓取 CNKI 网站上的相关文献。

4.1.2 爬虫优化

针对 CNKI 网站抓取文献过程中遇到的问题，进行如下优化处理:

(1)利用文献编号快速抓取文献。由于 CNKI 网站直接翻页无法实现，抓取的文献需要作如下处理:通过 Bio.Httpcrawl 的 Esearch 获取并存储文章编号;随后读取文献编号，通过 Bio.Httpcrawl 的 EFetch 抓取文献 。

(2)批量抓取文献提升抓取效率。文献抓取的过程包含以下 4 个步骤:向 CNKI 发送请求;在数据库中搜索结果;格式化成 XML 格式;将请求结果全部返回。

4.2 数据库设计实现

采用非关系型数据库 mongodb 存储爬取的文献结果，以及传统的关系型数据库 mysql 存储用于文献统计分析数据。爬取文献过程中有大量数据信息高并发频繁变更，文档型数据库 mongodb 以 bson 结构进行存储，对海量数据存储的读写速度比 mysql 有明显的优势。文献统计分析过程中，关系型数据库 mysql 在关联查询分析方面具备高性能。通过对平台信息的分析，利用表存储数据，数据库设计如下:

(1) 爬取阶段，mongodb 文献表 (article) 主要用来存放已经爬取的文献信息，mysql 爬取记录表 (crawlrecords) 主要用来做断点记录文献是否已经爬取。

(2) 数据分析阶段，表全部存放在 mysql 数据库中，mysql 文献表用来存储解析后的各种文献信息，如发表时间、关键词、期刊名、被引用数量等。

(3) 数据分析处理后，用于可视化的表也存在 mysql 中，发文量占比表 (proportion) 主要存储发文量、年限、占比等信息，用来展示国家发文比例变化趋势;期刊被引用量占比表 (proportion_of_journal) 主要存储期刊、年限、发文量、被引用量、占比等信息，用来统计期刊被引用数相对变化趋势;关键词次数表 (count_of_keyword) 主要存储国家、关键词、出现次数等信息，用来绘制关键词词云图;关键词占比表 (proportion_of_keyword) 主要存储国家、关键词、年限、占比等信息，用来统计热门关键词变化趋势。

4.3 可视化实现

主要利用 Ajax 技术向服务器发送请求，服务器收到请求后，读取相应数据库中用于可视化的数据，返回给 Ajax，用 Ajax 的 Success 方法对返回的 json 数据作相应处理，由 Echarts 渲染出可视化结果。

系统实现了用户可自由选择统计年限的功能。利用 Echarts 折线图、柱状图、堆叠区域图、平铺图、热力图对相应的统计结果进行可视化，并加入 Echarts 的工具栏，提供区域缩放、可视化结果保存的功能。

4.4 实现结果

4.4.1 系统环境需求

文献可视化系统的服务器硬件配置要求为: CPU: Intel Core i3 以上，内存 1 GB 以上，操作系统: Windows server 2012；环境配置要求: JDK1. 7. 0 及以上版本，PHP5. 5. 12，Apache2. 4. 9，MySQL5. 6. 17；用户浏览器要求: IE10. X 及以上版本 IE 内核浏览器、Firefox、Chrome 浏览器。

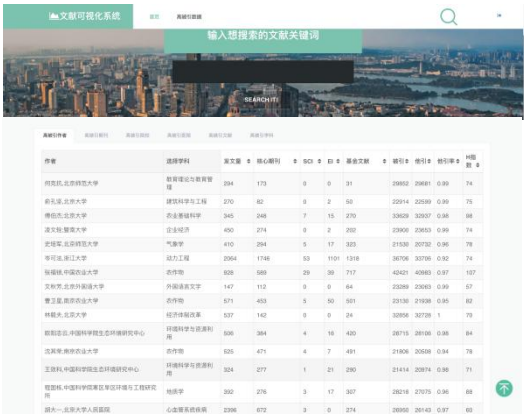


图 3 交互式文献可视化系统首页

4.4.2 可视化功能及应用

交互式文献可视化系统首页如图 3 所示，顶部菜单提供了高被引数据的功能，在页面下方即可浏览被引用量较高的文献信息。在首页的检索框中输入实体关键词后，类型抽取页面会生成关键词的结果说明，同时在说明下方生成该检索关键词的指数分析结果。本文以关键词“可视化”为例进行检索，最终生成结果如图 4 所示。

图 4 指数分析结果

指数分析图分别包括学术关注度、用户关注度、学术传播度、媒体传播度四部分。其中，每个生成的可视化结果均可进行不同类型的统计图展示，图 5 展示的是关键词“可视化”的折现图分析结果。

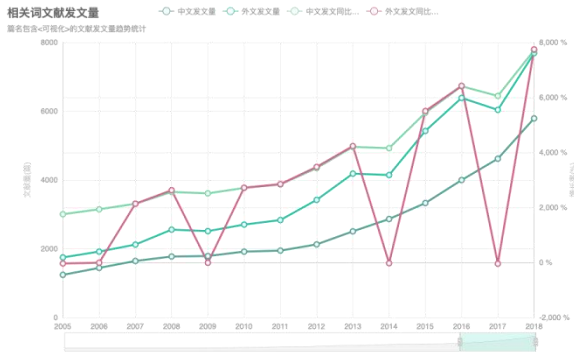


图 5 折线图分析结果

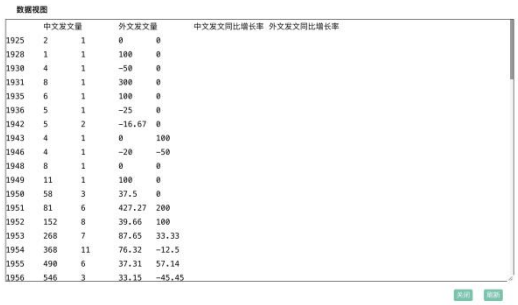


图 6 数据视图

用户在完成相关关键词的可视化分析后，可以查看并保存该结果的源数据，以便进行进一步分析，源数据信息如图 6 所示。

5.结语

本文设计实现了面向文献期刊数据库的交互式可视化系统，提供对文献数据的数据爬取、数据清理、实体识别、类型标签抽取、类型标注及知识实体关系图构建等功能，并通过 Web 数据可视化技术呈现给用户。同时，本文通过实验验证了本系统可以满足科研工作者进行初级文献可视化分析并了解某一学科领域的研究热点的需求，与传统方式相比具有一定的体验优化。因此，通过本系统可以简单便捷地获取到所关注学科领域的研究趋势、文献热点等，从而为科研工作者在科研方向上提供有价值的参考和启发。未来进一步的工作包括继续完善系统功能，提高系统后台处理性能，为用户提供更便捷、准确和高效的知识服务系统。

参考文献

[1]张学梅,汪伟歆.基于本体的期刊论文可视化检索系统研究[J].电子世界,2012(22):121-123.

[2]李信,程齐凯,刘兴帮.基于词汇功能识别的科研文献分析系统设计与实现[J].图书情报工作,2017,61(01):109-116.

[3]孟瑞丽. 基于复杂网络的江苏船舶产业集群研究[D].江苏科技大学,2012.

[4] BRADEN, R. A.; WALKER, A. D.Seeing ourselves: visualization in a social context. Readings from the Annual Conference of the International Visual Literacy Association (14th)[J].nternational Visual Literacy Association, Bloomington, IN, 1983 (2) :233-234.

[5]胡志刚, 侯海燕.科学技术学期刊群的可视化分析[J]大连理工大学学报 (社会科学版) , 2009 (2) :119-123.

[6]鲍杨, 朱庆华.近 10 年我国情报学研究领域主要作者和论文的可视化分析——基于社会网络分析方法的探讨[J]情报理论与实践, 2009 (4) :9-13.

[7] 张学福.信息检索可视化基本问题研究[J].中国图书馆学报, 2006 (3) :37-40

[8] 孙巍.基于引文的信息检索可视化系统研究[D].黑龙江大学信息管理学院, 2007:1-73.

[9] 陈颖.基于摘要信息的中文信息检索可视化系统研究与实现[D].黑龙江大学信息管理学院, 2007 (2) :1-56.