

信息检索可视化技术

李春旺

(中国科学院文献情报中心 北京 100080)

【摘要】 在讨论信息可视化功能级别、可视化基本技术、面向特定信息类型的可视化技术以及大型信息资源可视化技术之后,文章分析了三个具有代表性的可视化信息系统,最后阐述了当前信息检索可视化研究的主要任务以及今后的发展趋势。

【关键字】 信息检索 可视化技术 人机交互 **【分类号】** G354.49

Visualization Technologies for Information Retrieval

Li Chunwang

(Library of Chinese Academy of Sciences, Beijing 100080, China)

[Abstract] After discussing functional levels of the visualization of retrieved information, basic techniques, and visual techniques for special types of information and for large databases, the article analyzes three visual information systems, and put forward the main tasks today and a major trend.

[Keywords] information retrieval, visualization technology, human-machine interface

1 引言

信息检索可视化是将信息资源、用户提问、信息检索模型、检索过程以及检索结果中各种语义关系转换成图形,显示在一个二维、三维或多维的可视化空间中,帮助用户理解检索结果、把握检索方向,以提高信息检索的效率与性能。

可视化技术最早是针对科学计算提出的,目前的研究更多集中在信息可视化方面,其研究对象主要是大型数据库、网络资源以及各种类型信息集合^[1]。信息检索可视化的基础是计算机图形学和认知心理学,认知心理学揭示了人类感知和认识世界的方式,为信息检索提供可视化理论指导,计算机图形学提供了可视化实现工具。可视化技术不仅在揭示信息资源的广度与深度上有很大的优势,而且它能够将隐藏在信息资源内部的、复杂的、抽象的语义以直观的图形方式呈现给读者。面对不断膨胀的 Web 信息资源,可视化信息检索具有广泛的应用前景。

2 信息检索可视化技术

2.1 可视化功能分类

尽管信息可视化形式多样、纷繁复杂,但从功能上看,可视化技术可分为四个层面:信息空间、工作空间、认知工具、文献^[2]。信息空间是指研究信息资源、大型数据库及文献集的可视化,包括信息内容、存放地点以及信息结构等属性。由于不可能将这个空间中的所有信息一次性地可视化在一个有限的空间里,所以信息空间里的可视化研究又被分为针对不同资源类型的研究。工作空间是指可视化表现空间,分为二维、三维或多维空间,该层主要针对特定的信息对象和可视化需求,研究信息组织、存取以及空间显示方式等。认知工具是帮助用户理解信息的可视化工具,它应用特定的可视化技术实现一种或几种可视化功能,认

知工具通常为用户提供两种类型的可视化：静态图形、动态图形。文献是信息检索可视化的最小单元。该层主要研究文献内容、结构、属性的可视化问题。不同功能层面上的可视化研究产生了多种多样的可视化技术。

2.2 一般可视化技术

根据可视化对象、方式、目标的不同，可视化技术可划分为多种类型，其中最基本的可视化技术包括关联更新、广角与聚焦、聚焦加上下文、滤镜技术、空间显示、动态查询与过滤等。

(1) 关联更新技术 (Brushing and Linking)。当通过交互式方式更新一个窗口的信息设置时，关联窗口中相关内容将随之发生变化。这是可视化技术的一个最基本的思想。

(2) 广角与聚焦技术 (Panning and Zooming)。实现总体概要信息显示与个别具体信息显示之间的可视化切换。比如，文本聚类视图可以显示文献资源中与检索主题顶级相关的文献整体，也可以移动焦点到单篇文献上以及文献聚类上，便于揭示文献之间的关系，帮助读者理解检索结果。

(3) 聚焦加上下文技术 (Focus+Context)。放大聚焦结点，同时缩小周边对象，离聚焦结点越远的对象越小，不但可以突出重点信息，也能够揭示信息上下文关系。由于该技术具有鱼眼透镜的效果，因此又叫“鱼眼”可视化技术。

(4) 滤镜技术 (Magic Lenses/ Moveable Filter)。对透明窗口进行操作，通过在一个窗口上重叠不同效果的“滤镜” (不同的过滤条件与可视化算法)，实现信息的交互式选择、移动过滤、分析转换等功能，改变信息视图外形。

(5) 空间显示技术 (Starfield Display)。基于文献之间的相似度，将每篇文献都绘制成欧氏空间里的一个点或一片区域，相似的文献相邻，这些点构成的散列图揭示了文献内部以及文献之间的语义关系。

(6) 动态查询与过滤技术 (Dynamic Queries/Dynamic Query Filters)^[3]。B. Shneiderman 提出的动态提问与过滤技术目的在于直接操作信息资源，通过控制实时反馈、调整过滤参数、修订检索提问，过滤掉没有必要的信息，使用户能够直接聚焦在感兴趣的检索结果上。

2.3 面向特定信息类型的可视化技术

(1) 文献属性可视化。用 2D 方法显示文献信息，节点是文献，点的大小与颜色代表文献的不同属性，点的位置反应文献类别及相互关系，支持缩放、过滤和动态查询。比如，用 Starfield 技术可视化一个 HTML 文件时，点的大小表示文件的长短，点的颜色表示 HTTP 状态，可以支持 URL 请求和时间、URL 请求和访问地址的组合显示。

(2) 超链结构可视化。采用 2D 或 3D 技术动态显示网页间超链关系，每个节点代表一个网页，节点之间的有向边代表网页之间的链接，点、边的颜色代表网页与超链的有关属性。超链可视化将隐藏在网页内不可见的超链关系用图形的方式展示给用户，用户很容易判别“权威网页”和“HUB”网页。Websphinx^[4]系统的超链结构可视化具有一定的代表性，其超链图形随搜索引擎不断回传新的网页而动态增加节点，更新节点之间的超链接，并且标明网页状态 (死链、正在爬行节点等)。当鼠标移到节点上时，显示该节点的详细信息。

(3) 网站拓扑结构可视化。采用 2D 或 3D 方法来显示网站的拓扑结构，节点是网站，连线是网站之间的链接，节点的颜色代表网站的一种属性 (如访问频率)，用户可以通过鼠标拖放对图的布局进行调整。为准确表示网站的拓扑结构，需要消除无关的节点以及过多的

边,使图形更加简单明晰。VisVIP^[5]采用2D方法来显示网站的拓扑结构,它还可以将用户访问路径表示为有向图,直接绘制在网站结构图上,并以动画方式显示,通过联接访问序列中的相邻节点,可直接观看访问路径的拓扑结构。Astra SiteManager^[6]通过将两个不同时间上的两个网站合并,可以观察网站的变化。

(4) **日志信息可视化技术**。日志信息已经成为知识挖掘的重要资源。WebPath^[7]在一个3D的空间中绘制用户的浏览历史,节点是一个文档的代表,一个文档的多个代表用一根柱子串联,节点是一个立方体,表面的纹理是文档中的一张图片。当用户从一个文档访问到另外一个文档时,系统就绘制一个新的立方体,并用一条有向边连接开始立方体和新的立方体。垂直维固定为时间,其它两个维用户可以组合。系统还提供基于地理位置的布局。

2.4 面向大规模信息资源的可视化技术

在解决大规模信息资源可视化方面,XEROX PARC (Palo Alto Research Center)的研究比较引人注目。为了在有限的平面中显示更多的信息,XEROX PARC提出了双曲线浏览技术(Hyperbolic Browser)^[8,9,10]。这种技术在基于双曲线的圆形平面区域内显示层次结构信息,在相同的空间里,采用双曲线浏览技术显示的信息是普通2D技术的10倍。根在树的中间向外呈扇形扩展,当用户选择下游节点时,该节点被推到中心,同时放大其视图及细节信息。这种动态改变大小,使得用户在观察目标节点及其子节点时保持一个较大的上下文视图,能帮助用户从整体上理解数据资源,并通过提供感兴趣部分的细节信息,使用户可以对指定的信息进行深入的分析。双曲线树技术被用于开发浏览器、网站地图以及其它针对大型层次结构信息的可视化工具。特别适合浏览图库、文件系统、数据仓库、Web信息资源及其空间链接结构所包含的数据。

除双曲线浏览技术外,还有很多其它方案,包括Cone Trees^[11]以及NicheWork^[12]等。Cone Trees采用3D技术,将层次结构信息显示在一个虚拟空间中,每个节点和它的子节点布局呈锥形,为了让用户可以观察到所有的数据,锥形是半透明的,而且层次树可以转动。Cone Trees适合层次结构的信息可视化。NicheWork采用三种布局方式:圆、六角形和树,特别适合大型数据结构显示。

3 三个可视化系统分析

当前关于信息可视化的研究项目有很多,其中美国Maryland大学、IBM公司、HP公司的可视化研究具有一定的代表性。

3.1 FilmFinder^[13]

FilmFinder是美国Maryland大学设计开发的一个电影资料查询系统。该系统针对如何快速过滤结果集、如何渐进优化查询参数、如何用可视化方法分析检索结果等问题,提出了信息可视化检索的三个原则:

- (1) 动态方式提问,即允许方便、快速、增量式、可逆的查询参数变化。
- (2) 图形方式呈现,以二维的图形显示结果信息,提供可视化浏览功能。
- (3) 建立部件间的强关联,如检索参数发生变化视图随着变化、输出信息同时做为输入选项、节点与细节关联等。

FilmFinder系统在某种程度上模糊了“检索”与“浏览”的界限,试图利用可视化技术建立一种浏览式检索机制,它提出的几项原则对可视化信息检索的研究具有重要意义。

FilmFinder 也存在着以下问题:

- (1) 如何处理大型数据库系统及海量 Web 信息资源的问题;
- (2) 如何处理多种多样的信息类型的问题;
- (3) 如何解决查询条件扩大化的问题。

3.2 CoBrowse^[14]

CoBrowse 是 IBM Almaden 研究中心开发的一个基于传统浏览器的协作式浏览代理。它通过分析用户行为、检索词、浏览网页历史,推导用户的检索意图,预测用户感兴趣的网页,并以动态的、可视化接口实现人机之间的交互与协作。CoBrowse 采取 MAP (Multi-Attribute Pareto) 可视化方法,用二维图形组织 URLs,采取系统自动抽词与用户指定查询词相结合的方式,动态构建用户兴趣模型,实现建议信息及相邻信息的可视化。CoBrowse 通过实现检索与浏览之间的协作,试图将信息检索与信息浏览合并到一起。IBM 对 CoBrowse 的可用性进行了测试,结果表明,对于一个难度比较大的检索任务,使用 CoBrowse 能明显提高检索效率。

采用可视化技术协助浏览的信息系统还有: Atlas of Cyberspace^[15]、OLIVE-On-line Library of Information Visualization Environments^{[16][17]}等。

3.3 WAV 系统^[18]

WAV(Web Access Visualization)是 Hewlett-Packard Laboratories 开发的一个基于商业目的的可视化信息系统。它面向庞大的原始交易信息资源,利用可视化手段,对 Web 的存取行为、交易特征等属性进行研究,发现最有价值的客户以及相应的最好的服务策略等。WAV 根据相似度实现 Web 存取信息的可视化,为实现大型信息资源的非重叠可视化,WAV 提出了一种基于物理的弹性聚类引擎 (a physics-based mass-spring engine) 技术,并综合采用了层次化挖掘 (layered drill-down)、淡进淡出 (fade in/out)、多链接显示 (multiple linked views)、自动报警系统(automatic alarm system)等技术。

WAV 侧重于可视化技术在信息挖掘方面的应用研究,这是可视化技术在信息检索领域应用的一个重要方面。

4 结束语

当前,可视化信息检索研究的一个重要特点是将检索与浏览结合、将用户培训融合到检索过程之中。可视化检索认为,用户的需求不是一开始就明确、清晰的,检索过程中用户可能获得新的知识、非预期的检索方向,并随时调整检索需求与策略,经过反复检索最终获得所需求的文献^[19]。O'Day and Jeffries 研究发现^[20],80%的用户分析工作归结为6种类型:发现趋势、建立比较、汇总信息、确定关键子集、评价、解释。其余的20%包括交叉参考、概述、发现重要的可视化信息、其它混合行为。Richardo Baeza-Yates 等人将可视化信息检索过程划分为三个步骤:

- (1) 选择可视化开始点;
- (2) 确定图形提问表达方式;
- (3) 实现结果可视化呈现^[21]。

由于可视化技术可以很好地实现人机交互,是改进信息检索策略、提高信息检索效率的有效手段,它已经成为数字图书馆系统的主要特征之一。在今后的工作中,如何实现抽象

的、庞大的信息空间可视化仍是一个具有挑战性的问题，其中复杂对象结构、语义及相互间关系的可视化是重点，而建立在应用层面上的可视化信息检索与交互标准研究，是加速可视化技术应用进程的重要手段。

参考文献：

- [1] Tamara Munzner. Information Visualization.
<http://graphics.stanford.edu/~munzner/cga02/gei.pdf>
- [2] S K Card. Visualizing Retrieved Information: A Survey.
<http://www2.parc.com/istl/projects/uir/pubs/pdf/UIR-R-1996-01-Card-CGA-VisSurvey.pdf>
- [3] Shneiderman B. Dynamic Queries for Visual Information Seeking.
<http://citeseer.nj.nec.com/shneiderman94dynamic.html>
- [4] Websphinx. <http://www-2.cs.cmu.edu/~rcm/websphinx/>
- [5] VisVIP. <http://www.itl.nist.gov/iaui/vvrg/cugini/webmet/visvip/vv-home.html>
- [6] Astra SiteManager <http://tryandbuy.mercuryinteractive.com/>
- [7] Emmanuel Frécon, Gareth Smith. WebPath - A Three Dimensional Web History.
<http://davinci.informatik.uni-kl.de/vis98/archive/tp/papers/webpath.html>
- [8] John Lamping, Ramana Rao, and Peter Pirolli. A Focus+Context Technique Based on Hyperbolic Geometry For Visualizing Large Hierarchies.
<http://www.ramanarao.com/papers/startree-chi95.pdf>
- [9] Peter Pirolli, Stuart K Card, Mija M. Van Der Wege. Visual Information Foraging in a Focus + Context Visualization. <http://citeseer.nj.nec.com/pirolli01visual.html>
- [10] Peter Pirolli, Stuart K Card, Mija M. Van Der Wege. The Effect of Information Scent on Searching Information Visualizations of Large Tree Structures.
<http://portal.acm.org/citation.cfm?id=345304&coll=portal&dl=ACM&ret=1#Fulltext>
- [11] George G. Robertson, Jock D. Mackinlay, and Stuart K. Card. Cone Trees: Animated 3D Visualizations of Hierarchical Information
<http://www.cs.wmich.edu/~kkaugars/CS691/docs/p189-robertson.pdf>
- [12] NicheWorks. <http://willsfamily.org/gwills/NICHEguide/niche.html>
- [13] Ahlberg C. and Schneiderman B. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays.
<http://www.lri.fr/~mbl/ENS/DEA-IHM/papers/dynquery-chi94-ahlberg.pdf>
- [14] Denis Lalanne, Paul P. Maglio. interactive visualizations to facilitate human- machine collaboration in web browsing.
<http://diuf.unifr.ch/~lalanned/Articles/cobrowse.pdf>
- [15] Atlas of Cyberspace .http://www.cybergeography.org/atlas/info_landscapes.html
- [16] <http://www.public.iastate.edu/~CYBERSTACKS/BigPic.htm>
- [17] OLIVE-On-line:library of information Visualization Environments.
<http://www.otal.umd.edu/Olive/>

- [18] Ming C Hao, Pankaj Garg, Umeshwar Dayal, Vijay Machiraju, Daniel Cotting. Visualization of large web access data sets.
<http://www.hpl.hp.com/techreports/2002/HPL-2002-71.pdf>
- [19] Marcia J. Bates. The design of browsing and berrypicking techniques for the on-line search interface. *Online Review*, 1989(5):407-431
- [20] Vicki L. O'Day and Robin Jeffries. Orienteering in an information landscape: how information seekers get from here to there. In *Proc. Of the INTERCHI'93*. Amsterdam[Netherlands]: IOS Press .1993,4
- [21] Ricardo Baeza-Yates,Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley Longman, 1999
- [22] 黄如花, 张春蕾. 网络信息检索的发展趋势. *图书情报知识*, 2002,(4):48-50,66
- [23] 罗龙艳.基于可视化技术的信息检索初探.*现代图书情报技术*, 2002,(4):36-38, 30