信息可视化在数字图书馆中应用浅析

周静怡 孙 坦

（中国科学院研究生院 北京 100039） （中国科学院文献情报中心 北京 100080）

【摘要】 信息可视化是从科学可视化、数据可视化逐步发展起来的，信息可视化由结构化、显示空间化和人机交互界面三部分组成，将信息可视化技术引入到数字图书馆领域，解决信息需求与服务的个性化，信息提供的个性化等问题，可以通过信息可视化在一定程度上尝试解决其发展中遇到的问题，本文简单介绍了目前信息可视化在数字图书馆中的初步应用：信息检索过程可视化、信息检索结果可视化和映射知识领域可视化。

【关键词】 信息可视化 数字图书馆 检索结果 检索过程 映射知识领域 【分类号】 G250

Application of Information Visualization in Digital Library

Zhou Jingyi

(*Graduate School Chinese Academy of Science, Beijing 100039, China*)

Sun Tan

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

【**Abstract**】 Based on scientific visualization and data visualization, information visualization is developing. Information visualization consists of three parts: structuring, displaying and interface. Information visualization in digital library can solve some problems faced during its development, such as information requirement and service individuation, information supply individuation. This article simply introduces application of information visualization in information retrieval process, information retrieval result and knowledge domain in digital library.

【**Keywords**】 Information visualization Digital library Retrieval result Retrieval process Knowledge domain

1 信息可视化的基本概念

1.1 信息可视化发展过程

1987 年，美国国家自然科学基金会发表的一篇报告中，提出关于科学可视化概念。此后，科学可视化受到了极大的关注和广泛的研究，迅速发展成为一个新兴的学科。科学可视化（Scientific Visualization，SV）是从多个与计算机有关的学科中发展起来，其基本思想是“用图形和图像来表征数据”，将科学计算过程中及计算结果的数据转换成几何图形和图像信息显示出来并进行交互处理，成为发现和理解科学计算过程中各种现象的有力工具。

在科学可视化基础上，产生了现代的数据可视化（Data Visualization），即运用计算机图形学和图像处理技术，将数据转换为图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术。它涉及到计算机图形学、图像处理、计算机辅助设计、计算机视觉及人机交互技术等多个领域[3]。

收稿日期：2004－08－31

近年来，随着网络技术和计算机技术的发展，数据可视化概念已大大扩展，出现了信息可视化（Information Visualization），成为数据可视化新的热点。信息可视化是指将数据通过图形化、地理化形象真实地表现出来并且找出数据背后蕴含的信息。信息可视化相关技术能够实现对信息数据的分析和提取，然后以图形、图像、虚拟现实等易为人们所辨识的方式展现原始数据间的复杂关系、潜在信息以及发展趋势，以便能够更好地利用所掌握的信息资源[5]。

信息可视化作为一个研究领域出现后，在许多领域都引起了广泛的研究兴趣，其应用从股市动态图到最新的可视化授权的专利实验室，其目标是对来自于抽象数据的非视觉模式进行解释，最大的挑战是获取抽象和非视觉内容，将其转化为具体的、可触摸和可视的有意义的内容。从这个角度看，其定义如下：信息可视化是利用计算机支持的，互动的，将抽象数据增强认知的视觉表达。

1.2 信息可视化与科学可视化、知识发现和知识管理之间的关系

科学可视化是用图形和图像解释海量数据，而信息可视化是使用计算机支持的、人机交互的、视觉表现的方式，对抽象数据的认识的放大。信息可视化与科学可视化的主要区别首先是：科学可视化通常是观察基于物理的、有几何属性的数据，而信息可视化则用来显示各式各样的抽象数据；其次，科学可视化的用户多是高层次的专业工作者，而信息可视化的用户则主要是非技术人员；再次，信息可视化要为难以形象表达的抽象数据设计更加容易理解的表现形式，使其面临更大的挑战[5]。

数据挖掘与知识发现的许多活动都可以认为是一种可视化，即利用可视化技术进行信息传输、数据挖掘和知识发现，最终实现决策支持。从这种观点出发，数据挖掘和知识发现与信息可视化息息相关，在这些功能上，信息可视化与知识发现具有较多的重叠[3]。

知识管理是信息管理的高级阶段，信息可视化与知识管理互相渗透，互为依存。信息可视化借助于知识管理将信息对象进行综合、抽象、概念化、知识化，从而更方便简洁地实现可视化；信息可视化应用于知识管理（知识获取、知识组织、知识组织与服务）为用户提供了一个方便易用的知识环境。

2 信息可视化基本组成

信息可视化的对象只具有语义属性，任何对语义关系进行空间排序是作为信息可视化过程来完成的。从这个角度来说，信息可视化由以下三个过程组成。

2.1 结构化：将抽象的数据结构化

结构化，即模拟结构，其目的是为了区分信息基本关系和结构，通常使用的结构模型有表、树和网络。网络表达了现实世界和概念世界中更广泛的现实，如 Web，书目引文，一组图像等都可以被认为是一个网络。

2.2 显示空间化

提供一个可供用户交互和察看的显示空间，如二维空间或三维空间或一个曲状空间。信息可视化使用平面图和立体图来表达抽象结构已有很长一段时间的历史，目前在显示空间方法上已经有比较成熟的一些系统。

①Theme View 系统：该系统由美国太平洋西北国家实验室研发，利用改变词汇等级模式来查找典型的主题，使用户能在构建和可视化之间建立联系。

②VxInsight 系统：是由 Sandia 国家实验室（Sandia National Laboratory）开发的可视化系统，通过虚拟地图（Landscape）来模拟聚类信息，改编自目前流行的 Landscape 模型来显示潜在数据。该实验室的研究员运用该系统来显示 SCI 的聚类结构。

③自组织特征地图（Self－organized Feature Maps，SOMs）SOMs 在过去常用于信息检索中，ET－map 是 Yahoo！超过 100,000 与娱乐相关的网页的信息空间的多级类目的 SOM。

2.3 人机交互界面

人机交互界面指供用户与可视化描述进行交互的工具和方法，主要包括用户与空间显示的交互方式，用户可以改变空间可视的显示方式[1]。

3 信息可视化在数字图书馆中的应用

自美国科学家 90 年代初提出了数字图书馆概念后，以驱动多媒体海量数字信息组织与互联网应用问题各方面研究的技术领域开始在全球迅速发展起来。数字图书馆是以电子格式去存储海量的多媒体信息并对这些信息资源进行高效操作，如插入、删除、修改、检索、提供访问接口和信息保护等，其经过数十年的发展，已经进入一个相对稳定的阶段，同时面临多方面的挑战。将信息可视化技术引入到数字图书馆领域，解决信息需求与服务的个性化，信息提供的个性化等问题，可以通过信息可视化在一定程度上尝试解决其发展中遇到的问题。

3.1 信息检索过程可视化

信息检索由两个步骤组成：构建和使用。用户作为信息使用者的同时，也是信息构建者，通过增加检索路径到信息空间，这些增加的路径给其他用户检索其他路径提供了有价值的信息。语义空间中信息检索路径模型如图 1 所示。

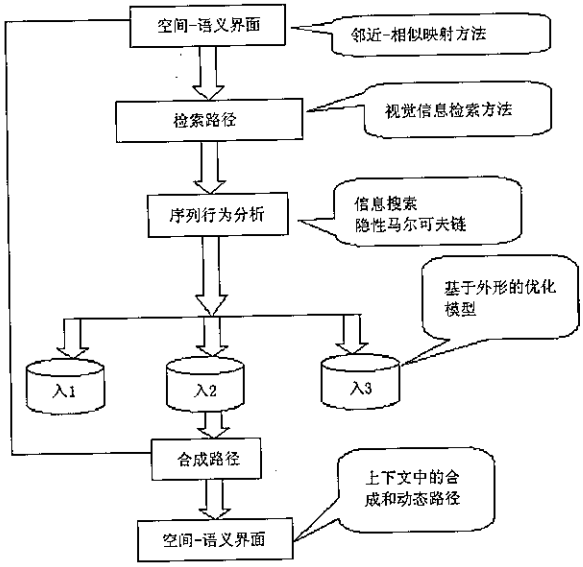

图 1 语义空间中信息检索路径模型

在该模型中，首先通过邻近－相似映射方法设计空间－语义界面，通过信息可视化分类法来对检索路径进行分析，主要是采用 Shneiderman 的信息可视化分类法，每个用户的行为根据检索过程进行精确测量，然后通过隐性马尔可夫模型（HMM）对信息检索行为进行序列行

为分析,在这个过程中包括网络寻址(Pathfinder Network Scaling)和多维尺度分析(Multidimensional Scaling,MDS)等可视化技术,成功的用户检索路径将插入隐性马尔可夫模型程序中。最后,通过隐性马尔可夫模型和在相应的语义空间中动态生成的合成用户路径。

3.2 信息检索结果可视化

目前,可视化技术已在工农业生产、交通运输、航空航天、科学研究和通信等领域得到初步利用。在信息管理中,信息检索结果可视化成果最为突出,已经取得许多成果。如马里兰大学人机互动实验室、Xerox PARC 关于数字图书馆研究、ISI 等都有相应的系统和软件。文中所列举的两个代表性研究成果都是来自于 Xerox PARC 关于数字图书馆研究的研究成果。

(1)Scatter/Gather

Scatter/Gather 系统是检索结果可视化中基于分类的文档簇法,通过使用一个活动的目录表来帮助用户掌握大量搜集的文档,首先系统使用文档聚集使搜集内容分解成小数量的连贯的文档簇,并将文档簇的简短概述表达给用户。基于这些概述,用户选择一个或更多的文档簇以供将来的学习使用。这些选择出的文档簇进行聚集或联合形成一个二次搜集。系统重新运用聚集技术来分散新的二次搜集以形成一个新的文档簇,并将这些依次表达给用户。在每一次连续重复下,这些文档簇将变得更小,因此也更具体化。

文档聚集算法充分运用速度来加强相互作用,而不是保证准确度。目前系统采用一次性聚集算法对文档搜集,采用连续性聚集算法进行预先处理文档搜集,一次性算法能够在一分钟以内在 SPARC20 工作站簇组织 500 次[4]。

它的主要思想是找出具有共词的文档,并把包含共词最多的文档放在同一簇中。每个簇根据簇中文档的主要语义内容给出一个总的标题,以便让用户能找到所需要的信息。当然,簇还只是完成了将文档进行归类的任务,为了揭示文档簇(集)之间的逻辑关系,还需要解决如何对簇进行排列。在簇的排列上,有的将簇作为结点排列成层次结构,有的排列成网状结构。检索结果如图 2 所示。

Cluster1 size:4 assistant director
603252"except service;consolidated listing of schedules A and B exception" 610814"5CFR Part737"
Cluster size217 section information 2 requirement regulation 3 request
690665"security is big business"
592791"organization;farm credit system financial assistance corp"

图 2 Scatter/Gather 显示图

(2)Tilebars

Tilebars 系统允许用户使用完全的信息,通过基于文档中查询检索词的分布式行为,决定哪些文档和文档的哪些部分来进行浏览,达到快速和简洁地显示:

①文档的相关长度;

②检索词在文档中出现的频率;

③检索词在文档中的分布和检索词与检索词之间的相对分布。

检索结果针对相关度最大的簇(从 Scatter/Gather 结果中得出),TileBar 被显示,横列反映了特定界面的标准:文档首先按在所有的检索词组中有多少页面数被命中来排序,然后是按检索词被命中的总数来排序,最后是通过相似的搜索来排序。显示的数字是原始的相似搜索顺序。

其具体显示方法是:用户在输入检索式时,将检索主题分成 n 组检索词(一般 n = 3),如检索词组 1,检索词组 2,检索词组 3,所有组的检索词都围绕一个相同的检索主题,根据每一组检索词构造文档簇。每一个大的矩形表示一个抽象的文档,矩形的长度表示文本的长度。每个矩形的上面的小正方形表示检索词组 1 的命中次数,中间的正方形为检索词组 2 的命中次数,下面的正方形为检索词组 3 的命中次数,正方形的颜色越深,表示检索词命中的频率越大(白色代表命中次数为 0,黑色为 8 或更多的命中次数,一个检索词组的所有检索词的频率为各检索词之和)。每个矩形的第一列为文档的第一部分,第二列为文档的第二部分,如此等等。用户需要浏览感兴趣的页面时,只需点击具体的正方形即可,而不必再为了查找具体的某一段而浏览整篇文档[4]。如图 3 所示。

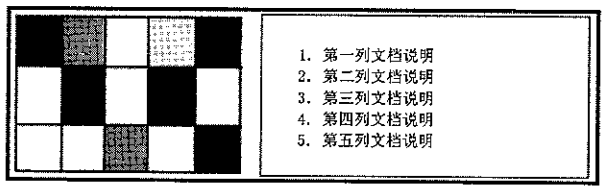

图 3 Tilebars 系统显示图

3.3 知识领域可视化

随着数字图书馆理论和实践的不断成熟,可供获取的知识也在不断地成几何指数增长,研究知识和科学的结构,科学发展的动力(如发展速度、发展模式等),寻找科学前沿等问题,越来越受到人们的重视。信息可视化技术的发展为研究知识领域的相关问题提供了新的思路,尤其是在最近的几年时间里,将信息可视化技术与科学计量学方法等相结合,生成具有各种属性的科学地图,表达学科、领域、专业、文献、著者之间的关系,解释知识领域的结构、映射知识领域的发展趋势,促进信息获取、使知识结构更加明显,帮助用户达到他们的目的。在国外有的学者将其称之为科学映射(Science Mapping)。

其中,文献同引(Co - citation)和共词(Co - word)是目前在研究知识领域可视化中,与信息可视化技术结合的比较成熟的两种方法,通过可视化软件生成地图来分析科学和技术发展的指标和方向。文献同引产生于 20 世纪 70 年代,随着《科学引文索引》的发展成熟, 它在科

学映射的历史上也具有独特的地位，主要有文献同引（DCA）和著者同引分析（ACA）等，相应的软件有 Pajek、Citespace、Histcite 等。共词分析方法产生于 20 世纪 80 年代，能够在不借助知识专家的帮助下提供客观的测量方法，通过聚类分析来观察高频词之间的亲疏关系，进而分析这些词所在的学科和主题的结构和变化。

4 结 语

信息可视化技术在许多领域已经开始发挥重要作用，如何发挥其在数字图书馆中的作用，将现有海量信息最大限度开发利用，解决信息需求与服务的个性化，信息提供的个性化等问题，仍然是一个值得探索和研究的问题。

参考文献：

1 Chaomei Chen. Mapping Scientific Frontiers: the Quest for Knowledge Visualization, Springer - verlag London limited 2003

2 周宁. 信息可视化在信息管理中的新进展. 现代图书情报技术, 2003(4):4 - 7

3 周海燕，郭建忠，王家耀. 知识发现与数据可视化技术浅析. 信息工程大学学报，2002，3(4):78 - 80

4 Marti Hearst, Gray Kopec, and Dan Brotsky. Research in Support of Digital Libraries at Xerox PARC. D_Lib Magazine, 1996 June

5 陈少强，走近信息可视化. http://media.ccidnet.com/media/ccu/567/02901.htm(Accessed Dec,2003)

6 周宁，文燕平. 检索结果的可视化研究. 中国图书馆学报，2002(6)

7 Gershon, N. D. 1994. (Panel chair.) Information visualization: The next frontier. ACM SIGGRAPH94 ConferenceProceedings, 485 - 486. New York: ACM

8 Sunishi Parikh. Visualization of Search Engine Query Result using Region based Dovument model on XML documents. 2000

9 HenrySmall. A SCI - MAP case study: building a map of AIDS research. Scientometrics, 1994, 30(1): 229 - 241

10 Eugene Garfield. Historiographic mapping of knowledge domains literature. Journal of Information Science, 2004, 30 (2): 119 - 145

11 Henry Small. Visualizing Science by Citation Mapping. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE. 50 (9): 799 - 813, 1999

（作者 E - mail：zhoujy@mail.las.ac.cn）

（上接第 4 页）

2 Kmi. OCML: Operational Conceptual Modeling Language. Summary. 2000 - 12. http://kmi.open.ac.uk/projects/ocml/, http://babage.dia.fi.upm.es/ontoweb/wp1/OntoRoadMap/show_lang.jsp?lang_name = OCML (Accessed Feb. 19, 2003)

3 Chaudhri, V., Farquhar, A., Fikes, R., Karp, P., and Rice, J., (1998). OKBC: A Programmatic Foundation for Knowledge Base Interoperability. In Proceedings 15th National Conference on Artificial Intelligence (AAAI - 98), pages 600 - 607

4 Cycorp, Inc. The CycL of Syntax. 2002 - 03 - 28. http://www.cyc.com/cycdoc/ref/cycl - syntax.html (Accessed Oct. 22, 2003)

5 Web - ontology working group. OWL Web ontology Language Overview. http://www.w3.org/TR/2003/PR - owl - features - 20031215/ (Accessed Feb. 02, 2004)

6 Dan Connolly, Frank van Harmelen, Ian Horrocks, et al. DAML + OIL Reference Description. 2001 - 03. http://www.w3.org/TR/daml + oil - reference (Accessed Nov. 30, 2002)

7 Mike Uschold. AIAI - TR - 192. Converting an Informal Ontology into Ontolingua: Some Experiences; A slightly abridged version of this paper appears in the Proceedings of the Workshop on Ontological Engineering held in conjunction with ECAI 96, Budapest; March 1996

8 AI/SRI. XOL: XML - based ontology - exchange language. 1999. http://www.ai.sri.com/pkarp/xol/ (Accessed Dec. 10, 2002)

9 The SHOE Team/CS/UMD. SHOE Simple HTML Ontology Extensions. 2002 - 04 - 28. http://www.cs.umd.edu/projects/plus/SHOE/ (Accessed Jan. 29, 2003)

10 Dan Brickley, R. V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Working Draft 23 January 2003. http://www.w3.org/TR/rdf - schema/ (Accessed Nov. 11, 2003)

11 OIL Steering Committee. Description of OIL - Ontology Inference layer. 1999 - 10. http://www.ontoknowledge.org/oil/ (Accessed Dec. 20, 2001)

12 ISI. Loom Project Hompage. 1999 - 07 - 12. http://www.isi.edu/isd/LOOM/LOOM - HOME.html (Accessed Nov. 11, 2002)

13 Michael Kifer and Georg Lausen. FLogic: A higher - order language for reasoning about objects, inheritance and scheme. In Clifford et al. [CLM89], pages 1341146

14 Pepper, Steve. 2002. The TAO of Topic Maps. XML Europe 2000. http://www.gca.org/papers/xmleurope2000/papers/s11 - 01.html (Accessed Aug. 25, 2003)

（作者 E - mail：lijing@mail.las.ac.cn）