

● 张学福

基于知识模型的文本信息检索可视化研究*

摘要 信息检索可视化是指把文献信息、用户提问、各种情报检索模型以及利用检索模型进行信息检索过程中不可见的内部语义关系转换成图形,在一个二维或三维的可视化空间中显示出来,并向用户提供信息检索的技术。基于知识模型的文本信息检索可视化,是利用信息资源的元数据信息来进行可视化检索。图1。参考文献29。

关键词 信息检索 可视化 知识模型 概念图

分类号 G354

ABSTRACT In this paper, the author introduces the definition of information retrieval visualization. The text information retrieval visualization based on knowledge models is the visualized retrieval with metadata information of information resources. 1 fig. 29 refs.

KEY WORDS Information retrieval. Visualization. Knowledge model. Concept map.

CLASS NUMBER G354

20世纪90年代中期开始,信息可视化技术得到迅速发展,信息可视化的成果不断地应用于信息检索。信息可视化在信息检索中的应用为现有信息检索问题的解决在一定程度上提供了帮助;另一方面,知识领域概念图从20世纪80年代起在知识的组织、表示和共享方面得到应用。基于概念图的知识模型在90年代初提出并逐步得到应用。本文希望通过把知识模型描述机制与信息检索可视化结合起来,在满足用户深层次信息需求方面做些探索。

1 概念图、知识模型和信息检索可视化

概念图(Concept maps)是一种提供可视化信息表示的方法,它利用人类的视觉能力来理解复杂的信息^[1]。1993年,Novak创建了一个派生于学习理论的方法——概念映射(concept mapping),用它来描述由链接和结点组成的网络里的概念及它们之间的关系,用结点描述概念,链接描述关系。链接可以被标注,可以是无方向、单方向和双向的,并且能显示概念之间暂时或偶然的关系。该方法能够鼓励用户在已有的认知结构基础上吸收新概念和主题。概念图能被用来定义新的概念,交流复杂的思想、明确集成新知识和原有知识来辅助学习^[2]。

知识模型是指关于特定知识领域的概念图及其相关资源的集合。这里资源是与概念图相联接的文件,它可以是任何形式的,包括视频、图像、文本、网页等或其他的概念图。知识模型能帮助解释和补充概

念图中的信息。Cmap Tools是人机识别研究院IHMC(The Institute of Human & Machine Cognition)开发的软件,利用它,用户能够创建、导航、共享和分析评价以概念图形式表示的知识模型,它具有简单和便于共享的特性,一般专家学者能够直接构建其相关领域的知识模型。

信息检索可视化是指把文献信息、用户提问、各种情报检索模型以及利用检索模型进行信息检索的过程中不可见的内部语义关系转换成图形,在一个二维或三维的可视化空间中显示出来,并向用户提供信息检索的技术。

为了简化研究问题,把信息检索可视化对象限定为文本信息。也即基于知识模型的文本信息检索可视化,它是利用信息资源的元数据信息来进行可视化检索,在此基础上,集成基于概念图的知识模型描述机制,提供知识模型内的可视化浏览和检索。

2 文本信息检索可视化研究现状

2.1 国外信息检索可视化研究现状

信息检索的可视化研究可以追溯到20世纪60年代。由于信息检索可视化应用面临很多不易解决的问题,现在的系统大多停留在原型系统的研究上。国外在信息检索可视化研究和应用方面的进展情况如下:

在内容描述方面采用共频现象分析(Co-occurrence Analysis)。共频现象分析的基本出发点是,如

* 本文是黑龙江大学杰出青年基金项目“现代信息检索理论与应用研究”成果之一。

果两个词经常在同一篇文章中共同出现,这两个词之间就一定有一些关系。若两篇(或多篇)科学文献有一个(或多个)相同的词,则这两篇(或多篇)文献或其相应著者间必然存在一种潜在的关系^[3]。Chen, H. 等用其指导语义索引和由机器生成主题词表^[4]。Lin 等在 AuthorLink、ConceptLink 和 PNASLink 原型系统中用它作为资源描述的方式^[5]。

可视化算法主要有 Kohonen's feature map (SOM) 和 Pathfinder Network (PFNETs)。SOM 在信息检索方面的应用主要有:Lin 等使用它来构建信息检索的自组织语义图,语义图能将输入文档间的语义关系可视化^[6~7]。Chen 和 Carr 在一些项目中应用 PFNETs,把用户的认知图和文献检索连接起来;Lin 等在 AuthorLink、ConceptLink 和 PNASLink 原型系统中,使用了 PFNETs 映射^[8~9]。

可视化映射形式包括全局映射和局部映射。全局映射与局部映射均以共频现象分析为基础。Vxinsight、Galaxies 等采用全局映射^[10~11],而 AuthorLink、ConceptLink 和 PNASLink 原型系统采用局部映射。

McCain 概述了传统 ACA 的典型流程^[12]。

Shneiderman 提出如下可视信息查询原则:先整体浏览、缩放和过滤、然后根据需要选择细节^[13]。

代表性的交互模型有两个^[14]。Foley 等提出一个四层次方法:概念层、语义层、句法层和词汇层,用于设计和评价交互式系统。Marchionini 定义了一个强调信息的人类查询模型,它包括用户、任务、域、系统、结果和设置等因素。

White 和 McCain 提出 6 项可视化信息检索接口评估标准^[15]。

信息检索可视化原型系统主要有 AuthorLink、ConceptLink、PNASLink 和 Citespace^[16~19]。AuthorLink 系统是基于作者共引分析(ACA)的,被用来分析 20 多年来的科学和学术知识结构。ConceptLink 是基于词共生频率创建的,用于表示可视化概念之间的关系。它采用了两种映射方法,PFNET 映射和 Kohonen 映射,前者显示词之间最相关的链接和联系,后者提供了一个更好的浏览。PNASLink 系统的特征是在一个系统中集成多种类型的索引和多种视图。除了包括 AuthorLink 和 ConceptLink 的特征,也能被映射作者和关键词,以及杂志之间的关系。Citespace 是由陈朝美博士开发的,用于分析和可视化文献资源关系的系统。

2.2 国内信息检索可视化研究现状

信息检索可视化研究基本处于对国外研究的跟踪阶段,主要的研究单位有武汉大学信息资源研究中心(承担教育部重点课题“信息可视化与知识检索”)

刘玮等介绍了信息可视化两方面的应用。在文献资源可视化的基础上,给用户提供了检索操作过程的可视化和检索结果的可视化^[20]。

罗龙艳比较了基于可视化技术的信息检索与传统检索的优劣,着重阐述了可视化信息检索的基本原理及过程^[21]。

李君君介绍了 Crestani 和 V. Rijssbergen 根据映射概念建构的信息检索逻辑模型^[22]。

周宁、文燕平介绍了检索结果可视化的常用方法^[23]。周静怡、孙坦介绍了信息检索过程、检索结果的可视化^[24]。

李爱国、汪社教、李春旺介绍了信息检索可视化技术^[25~26]。

文燕平认为 WWW 信息检索可视化实现包括 4 个步骤:第一,确定可视化对象,并根据可视化对象间的关系抽象出虚拟结构,方便用户找到相关的检索结果;第二,将抽象信息特征映射成空间化、图形化的特征;第三,根据映射结果,选用合适的可视化的隐喻形式,对可视化对象进行组织,构建可视化空间;第四,绘制可视化图形^[27]。

文燕平提出一个多层次的信息检索可视化概念模型。它可以根据用户不同的认知层次、不同的检索需求、检索结果所属的不同领域等因素,抽取不同形式的可视化对象,形成不同的可视化数据模型,生成不同的可视化空间,提供不同的可视化显示形式^[28]。

2.3 信息检索可视化技术分析

2.3.1 人工参与的非实时可视化检索

这里以传统 ACA 的典型流程为例,其流程为:(1)选择作者;(2)检索共引频率;(3)编辑原始的共引矩阵;(4)把共引矩阵转换为一个相关矩阵;(5)相关矩阵的多元分析(使用要素组件分析、聚类分析和多维测量)。

其优点是词语选择经过领域专家参与,词语选择比较规范,可视化效果较好。不足是属于全局可视化,实时交互困难;流程中包括许多人工处理过程及多个不同的计算机处理的系统,难以实现动态实时处理,在可视化检索发展过程中处于过渡阶段。

2.3.2 动态实时可视化

动态实时可视化的流程为:(1)选择数据资源;(2)选择关联词语;(3)建立关联词语共频次数原始矩阵;(4)将原始矩阵转换为相关矩阵;(5)用SOM、PFNETs等算法进行可视化映射;(6)视图绘制。

优点是:直接以词语共频矩阵的行作为输入,进行自动聚类;能在一定程度上反映词语之间的关系;可对实际数据库中的数据进行处理;可以实现动态实时可视化检索,是可视化信息检索的发展方向。不足是:(1)现有的SOM、PFNETs映射图在一定程度上反映了词语之间的关系,但无法反映出词语之间是何种关系。(2)现有原型系统一般采用受控词标引,虽然提供电子形式的词表(如MeSH),用户使用还是不太方便。另外,在使用非受控词时可能会出现问题。(3)现有的原型系统还不能为用户提供类似基于概念图的知识模型的视图,在其视图中提供某一学科或分支学科的较多信息时有一定困难。

3 基于知识模型的文本信息检索可视化系统分析

3.1 必要性

是用户在数字化科研资源环境中交流互动的需要。用户希望检索系统能够以新的视角为他们提供信息检索,由其控制检索过程,并且将现有检索系统中不可见的内部操作过程可视化。用户可以根据需要以人机互动方式在查询视图中进行选择,缩放视图,分析视图中关联词语之间的关系,选择获取文献信息。

是满足用户获得某领域资源整体浏览图的需要。在论著写作过程中,一般要求对文献进行综述,这就需要从相应资源整合上进行把握,包括研究对象最早研究的时间、研究人员,现有研究的分支学科,各研究领域的主要作者、主要的研究项目、主要的研究机构、主要期刊等等;这种浏览还能帮助用户在作更深入研究前查看一个新的领域。如果用户在一个新的专业领域开始研究,浏览能有助于他发现有哪些其他的研究领域被涉及到,以及它们之间的关系,某研究领域的研究情况等等。在整体资源浏览图的指导下可以较好地完成任务,并能节省用户时间。

是满足用户构建其知识结构的需要。用户在科研过程中通过学习研究,充实完善了自己的知识结构,积累了大量资料。他们迫切需要有效地梳理这些资源,如果能够按照知识结构的角度梳理则能方便识别出在知识结构组成部分都有哪些知识储备,并根据其具体需求,使用和补充。

3.2 需求分析

现行的基于布尔逻辑模型的检索系统提供浏览和关键词检索两种途径。用户针对“问题”在该系统中查找资料,浏览(分类、主题描述)和关键词检索两种途径的结果是按时间顺序、字母顺序,或者是文献与查询之间的一些权值顺序来显示检索结果。如果用户要知道该领域研究的最早时间、研究人员,现有研究的分支学科,各研究领域的主要作者、主要的研究项目、主要的研究机构、主要期刊等等,就要逐一阅读,然后归类整理;如果文献数量过大,可能还需要请专业人员帮助整理。林夏博士等研究开发的信息检索可视化原型系统(ConceptLink等),从作者、关键词、期刊等角度揭示了这些关联词语之间的关系,用户通过输入关键词可以获得与该词语相关的词语关联视图。但关联视图采用无向图表示,在揭示是什么样的具体关系时还存在不足,因此要获得前述需求,用户要做很多工作。用户可能需要把这些信息按照知识结构保存下来,这些活动的处理还需要进行探索。

用户在科学研究过程中,通过查询活动获得的信息仅是其研究活动的基础。检索系统应该尽可能地帮助用户缩短查找、分析信息的时间,由系统界面完成和用户的交互,交互过程的内部实现用户信息查询行为的代理,通过对现有检索系统的改进,满足用户科研交流互动的需求,以知识结构的形式呈现浏览结果等。

4 基于知识模型的文本信息检索可视化系统设计

4.1 系统设计的原则

(1)关联词语采用自然语词(关联词语是指经常共频的词语),共现频率排序越高的词语在映射中的价值越大。

(2)用户认知负担最小原则,用户通过输入单一的词语(种子)来生成相关映射图。

(3)映射程序和搜索引擎紧密结合,充分发挥现有搜索引擎的功能。

(4)从完整的元数据数据库中,而不是从特定元数据集合中进行检索。

(5)用户能跨视图域查找种子词语的词语关联,即不仅能在本类型视图中找到,还能在不同类型的视

图中找到。例如:系统应该能够把种子作者转换为关联关键词或把种子关键词转换成关联作者等。

(6)词语关联共引矩阵实时生成。

(7)映射过程在秒级完成,最好在数秒内。

(8)映射图能够揭示元数据数据库中词语之间的词语关联。

(9)采用 PFNETs 映射算法,在二维空间中展现共现频率排序高的词语之间的关联。

(10)可视化处理过程与数据库大小关联程度最低。

(11)较强的交互性用户能够根据需要在多种视图中控制屏幕对象,获取信息^[29]。

4.2 系统的框架结构

体系结构如图 1 所示。它包括三个层次:底层数据层,主要包括信息资源数据库、共现矩阵数据库和搜索引擎;中间层,主要包括应用服务器、各种映射算法等,实现与用户界面、数据库的交互,各种可视化映射等;基于知识模型的用户界面,该层主要实现二维实时信息检索可视化和基于概念图的知识模型信息的浏览。

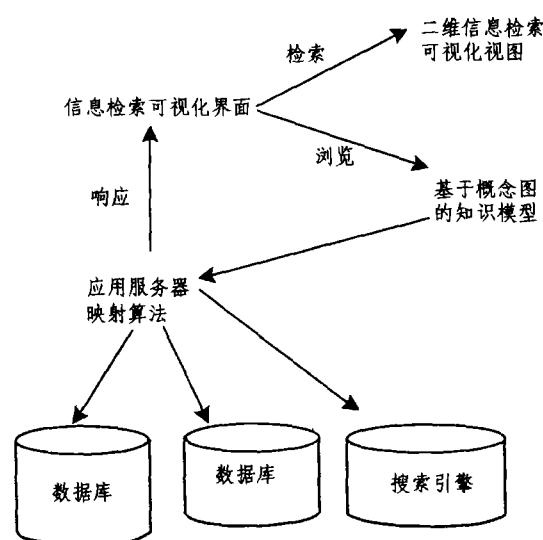


图 1 基于知识模型的信息检索可视化模型体系结构

该系统的功能模块主要以下 6 个。自动过滤子系统,主要是从数据库中把不含种子词语的文档滤掉。关联词语选择、统计子系统,主要是对词语统计,选择关联词语。关联词语共频矩阵生成子系统,主要是根据关联词语原始矩阵生成共频矩阵。关联词语映射、映射图绘制子系统,主要是利用共频矩阵数据,

按照 PFNETs 算法生成映射图。知识模型构建子系统,主要是通过手工构建某领域知识模型。知识模型节点链接自动生成和修改子系统,主要是根据实时可视化检索结果,对知识模型的部分内容进行自动修改。

参考文献

- 1 Concept Maps. <http://classes.aces.uiuc.edu/ACES100/Mind/CMap.html> (2005-07-21 查询)
- 2,3,5,14,18 林夏. 信息可视化与内容描述. 现代图书情报技术, 2004(10)
- 4 Chen, H., Schatz, B., Yim, T., & Fye, D. Automatic Thesaurus Generation for an Electronic Community System. Journal of the American Society for Information Science, 1995, 46(3)
- 5 Lin, X., Soergel, D., & Marchionini, G. A self-organizing semantic map for information retrieval. Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, 1991, 262 - 269
- 6 Jan W. Buzydowski, Howard D. White, Xia Lin. Term Co-occurrence Analysis as an Interface for Digital Libraries. Lecture Notes in Computer Science. 2002, Volume 2539, 133-144
- 7 Chen, C., & Carr, L. Trailblazing the Literature of Hypertext: An Author co-Citation Analysis 1989-1998. Hypertext'99, Proceedings of the 10th ACM Conference on Hypertext, Darmstadt, Germany, 1999, 51-60
- 8 Chen, C., & Carr, L. Visualizing the Evolution of a Subject Domain: A Case Study. Proceedings of IEEE Visualization 99, San Francisco, California, USA. 1999, 499 - 502
- 9 http://www.cs.sandia.gov/projects/VxInsight.html (2005-07-15 查询)
- 10 http://www.seds.org/messier/galaxy.html (2005-07-15 查询)
- 11 McCain, K. W. Mapping authors in intellectual space: A technical overview. Journal of the American Society for Information Science, 1990, 41(6)
- 12 Schneiderman, B. The Eyes Have it: a Task by Data Type Taxonomy for Information Visualizations. Proceedings of Symposium on Visual Languages, Boulder, CO, Proceedings of IEEE, 1996, 336 - 343
- 13 White, H. D., McCain, K. W. Visualization of Literatures. Annual Review of Information Science and Technology, 1997, 32:99. 168
- 14 http://project.cis.drexel.edu/authorlink/ (2005-03-15 查询)

- 查询)
17 <http://cite.cis.drexel.edu/#ConceptMap>(2005-03-15 26 李春旺. 信息检索可视化技术. 现代图书情报技术, 2003
查询)
(6)
19 <http://cluster.cis.drexel.edu/~cchen/citespace/>(2005 27 文燕平. WWW信息检索可视化实现原理研究. 现代图
-08-15查询)
20 刘玮等. 基于文本的信息可视化方法研究. 现代图书情 28 文燕平. WWW信息检索可视化研究. 武汉大学博士论
报技术, 2003(2)
文, 2004
21 罗龙艳. 基于可视化技术的信息检索初探. 现代图书情 29 Lin, X. ; White, H. D. ; & Buzydowski, J. Real-time author
报技术, 2002(4)
co-citation mapping for online searching. International Journal
22 李君君. 基于映射的信息检索逻辑模型. 情报理论与实
践, 2004, 27(4)
23 周宁, 文燕平. 检索结果的可视化研究. 中国图书馆学
报, 2002(6)
24 周静怡, 孙坦. 信息可视化在数字图书馆中应用浅析.
现代图书情报技术, 2005(1)
25 李爱国, 汪社教. 信息检索可视化. 现代图书情报技术,
~~~~~  
(上接第46页)

基于数据库的知识发现的关键,集中在与数据获取任务相关的问题。数据选择的错误将严重影响整个知识发现过程,导致失败。为了增强用户关于地理数据有效性的意识,INVISIP 提出了一种基于可视化数据挖掘的知识发现方法。它允许用户完成确定的和探索性的分析,帮助他在所需要的数据和可用的数据之间找到折中。这样,通过使用可视化作为交流渠道,即使是一名不熟练的用户,也能较好地发现自己所需的知识。

#### 参考文献

- 1 Badjio, E. F., Poulet, F. Dimension Reduction for Visual Data Mining. <http://asmda2005.enst-bretagne.fr/IMG/pdf/proceedings/266.pdf> (2005-12-10 查询)
- 2 Daniel A. K. Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics, 2002(1)
- 3, 9, 11 Edward J. W. Visual Data Mining. <http://www.galaxy.gmu.edu/stats/syllabi/infi979/VisualDataMining.pdf> (2005-12-10 查询)
- 4 Pak Chung Wong. Visual Data Mining. [http://www.pnlgov/infviz/visual\\_data\\_mining.pdf](http://www.pnlgov/infviz/visual_data_mining.pdf) (2005-12-12 查询)
- 5 刘绪崇. 基于OLAM的可视化数据挖掘技术研究. [学位
- 2004(2)  
26 李春旺. 信息检索可视化技术. 现代图书情报技术, 2003  
(6)  
27 文燕平. WWW信息检索可视化实现原理研究. 现代图  
书情报技术, 2005(4)  
28 文燕平. WWW信息检索可视化研究. 武汉大学博士论  
文, 2004  
29 Lin, X. ; White, H. D. ; & Buzydowski, J. Real-time author  
co-citation mapping for online searching. International Journal  
of Information Processing & Management, 2003, 39(5)

张学福 博士, 黑龙江大学信息资源管理研究中心教授, 黑龙江大学信息管理学院副院长, 教授。通信地址: 哈尔滨黑龙江大学信息管理学院。邮编 150080。

(来稿时间: 2005-12-05)

论文]. 长沙: 国防科学技术大学, 2002

- 6 刘勤等. 基于平行坐标法的可视数据挖掘. 计算机工程与应用, 2003(5)
- 7 欧海英等. 平行坐标可视化技术在固体火箭发动机优化设计中的应用. 固体火箭技术, 2004(4)
- 8 Jing Y. , Matthew O. W. , etc. Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets. <http://davis.wpi.edu/~xmdv/docs/vhdr.pdf> (2005-12-14 查询)
- 10 Dianne, C. Calibrate your Eyes to Recognize High-Dimensional Shapes from Their Low-Dimensional Projections. <http://www.jstatsoft.org/v02/i06/paper.html> (2005-12-21 查询)
- 12, 13, 14 Albertoni, R. , et al. Knowledge Extraction by Visual Data Mining of Metadata in Site Planning. <http://www.ima.ge.cnr.it/irma/personal/albertoni/PersonalPage/src/SeanGIS2003.pdf> (2005-12-21 查询)

余肖生 武汉大学信息管理学院2004级博士研究生。  
通信地址: 湖北武汉。邮编430072。

周宁 武汉大学信息管理学院教授。通信地址同上。  
张芳芳 武汉大学信息管理学院2004级博士研究生。  
通信地址同上。

(来稿时间: 2006-01-10)