●周 宁 文燕平

检索结果的可视化研究*

摘 要 将检索结果用图形进行可视化显示,是改善检索效率的有效途径。信息可视化的三要素是:空间化、显示空间、用户交互机制。检索结果可视化的常用方法为:基本分类的文档簇法、基于超链接法、基于语义内容法。图 2。参考文献 12。

关键词 检索结果 信息可视化 WWW 信息检索
分类号 G252.7

ABSTRACT The visualized graphic display of search results is an effective approach to improve retrieval efficiency. The three basic elements of information visualization are spatialization, display space and user interactivity. In the paper, the authors also propose some usual methods. 2 figs. 12 refs.

KEY WORDS Search results. Information visualization. WWW. Information retrieval.
CLASS NUMBER G252.7

1 引言

因特网信息的爆炸式增长引起了对信息过滤及信息组织的关注,其目的在于实现高效准确的信息检索。由于因特网信息空间大、数据的多样性以及缺乏通用的索引机制,使得人们在因特网上搜索信息感到棘手不便。目前用得最多的搜索工具有两类:一类是基于 spider 的自动搜索系统,如 Lycos、Harvest 等;另一类是人工组织系统,如 Yahoo 和 Internet Yellow pages 等。这些工具在检索结果的显示上一般都采用列表的方式,按照相关度的大小顺序将文档的标题、URL、关键词、摘要等基本信息提供给用户,用户根据所提供的 URL 进一步获取全文。这种一维的线性排列仅仅是完成了结果的提供任务,它在对检索结果之间的关系的揭示、与用户的交互等方面则无能为力。用户可能在列表中找到了一篇满意的文档,但若想从这篇文档进而找出与它相关的文档时,不得不从头进行一次新的检索。此外,线性排列的方式对大文档集只能分屏显示,用户难以从总体上了解整个结果之间的关系。用户必须逐篇扫描列表上的文档,才能判断出哪篇文档不相关,哪篇文档需进一步获取全文。尽管检索结果是按照相关度依次排列,有用的文档一般都排在前面,但真正符合用户需要的文档可能排在最后面。而用户往往很难有时间和耐心把整个列表扫描完。而且各个搜索引擎的排序依据是根据他们自己的关键词的权重进行计算的,这与用户的理解不可避免地存在偏差。

因此,如何显示检索结果,帮助用户快速获取所需的信息,具有很大的实用价值。人类认知和理解事物的一个特征是利用图形。使用图形来表示信息,可以赋予信息某种虚拟的形态,目的是辅助人们分析、综合信息及其信息之间的关系,减少理解和认知它们所需的努力[1]。从人类认知的角度出发,设计和创建各种信息可视化工具来表示检索结果,是改善目前网络信息检索的一种有效途径。将检索结果用图形化可视化方式进行显示不仅可以使人们直接观察到信息,也能实现与用户更直接、直观的交互;不仅能揭示检索结果中文档之间的关系,还能揭示出检索词与所检索到的文档之间的关系。与传统的滚动列表相比,用户不仅能从中快速找到符合要求的文档,也能对所检索的主题获得全面了解。此外,可视化的特征如颜色、位置等信息能帮助用户快速找到感兴趣的区域。

2 信息可视化的三要素

可视化(Visualization),《牛津英语词典》解释为:"构成头脑情景的能力或过程,或不可直接察觉的某种东西的视觉。"Haber 和 McNabb 指出,可视化是"一系列的转换,这种转换将近原始模拟数据转换成可显示的图像。这种转换的目的,在于将信息转换

*本文系教育部重点项目《信息可视化与知识检索》的子课题研究成果。

成可被人类感应系统所领悟的格式。"[2]其实,可视化不仅是指变不可见为可见,更重要的是指各种不同目的思维过程中的可视化分析,这是指利用可视化去探究概念及作为概念加工和深化的一条途径。

与科学可视化一样,信息可视化也是通过一种可视化的表现形式来帮助人们进行理解。但信息可视化还面临一些科学可视化所没有遇到的难题。对于科学可视化来说,可视化描述是从一种实际存在的现象所得到的,因此,根据该现象的空间、时间及其他属性就能对该现象进行可视化描述。而对信息可视化来说,要可视化的对象并不具备那些属性,只具有语义属性,任何对语义关系进行空间排序也是作为信息可视化过程来完成的[3]。从这个角度来说,将语义信息空间可视化必须解决:①将抽象的数据空间化,包括两方面:数据组织和数据的可视化空间描述;②提供一个可供用户交互和察看的显示空间,如二维或三维空间或一个曲状空间;③供用户与可视化描述进行交互的工具和方法。信息可视化的这三个要素缺一不可。各部分具体内容见图1。

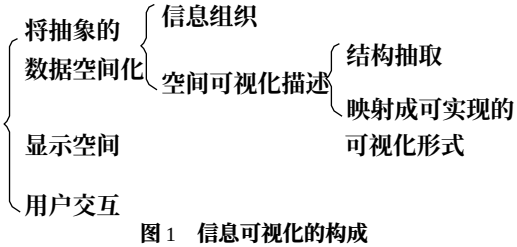

图1　信息可视化的构成

第一要素:空间化。空间化包含两方面的内容,一个方面是组织信息或数据的过程,而信息或数据是要被可视化的。例如,抽取出关键词构造一个多维的向量空间。在有些系统中,要显示的数据已经组织好,比如文件系统。空间化的另一方面是要创建一个可以观察到的、可显示的、空间可视的描述。最终的空间属性是由这一步得到的。

第二要素:显示空间。如上文所讲的,显示空间是显示给用户的可视化空间。

第三要素:用户交互机制。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。

在解决显示空间和用户交互时,必须遵循人机交互的两个基本原则:可见性(Visibility)和可伺服性(Affordance),即所有的对象都是可见的,并且控制(操作和操纵)与控制结果间存在良好的映射。一个控制严格对应一个功能,具有良好的反馈,即用户的目标、需要的动作以及结果都是可感知的、有意义的并且不是随意规定的,这样便于用户的理解[4]。"可见性"是指所有的对象本身应该对用户可见并且用户对对象的操作过程也应该可见;"可伺服性"是指对象的属性,以及该特定对象所提供的操作和操纵是可以感知的,也就是说,界面提供的实际内容应该与目标用户想象中该对象包含的属性和操作尽可能地吻合。

3　检索结果可视化的主要方法

3.1　基于分类的文档簇法

要在一个有限的显示空间中将所有检索结果显示出来,必须在结果显示上确立一个合理的逻辑结构。目前普遍采用的策略是构造文档簇,它的主要思想是找出具有相同词的文档,并把包含共同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,为了揭示文档集(簇)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。该方法的典型代表是 Scatter/ Gather 系统。

3.2　基于超链接法

利用文档之间的超链接将检索结果文档之间的关系可视化是最直接、最省力的方法,它可以为人们进一步扩展浏览 WWW 文档信息提供导航。超链接不仅指明文档的逻辑结构,也具有和用户交互等重要扩充功能。采用这种方法的代表有:McCahill and Erickson 针对 Gopher 提出了三维空间接口,将 Gopher 中文件结构或检索结果中的文档用各种形状和纹理的图标来表示[5]。Harmony 浏览器就一篇文档所有的链用二维的结构地图表示,用三维的风景图表示链的结构层次[6]。Narcissus 系统把 WWW 中的链之间的关联三维可视化表示,并根据用户对文档的操作决定文档在三维图的位置[7]。

根据由作者提供的超链来建立可视化文档之间的联系具有容易实现的优点,但是 HTML 语言仅能描述文档间的交叉链接关系,而不能支持层次链接关系;而且检索结果文档往往来自不同的站点,这种广域分布特性难以通过超链接支持基于全局结构的 WWW 文档的组织和浏览。由于 HTML 文档的上述

缺陷,降低了 WWW 文档的连贯性,使得用户无法基于良好的全局结构进行检索结果的组织和浏览,容易产生"碎片"感,从而影响文档的浏览效率,导致"迷航"。而且用户真正关心的是文档的内容之间的联系,而链结构只能部分地反映出这种内容之间的联系。因此出现了基于语义内容法。

3.3 基于语义内容法

目前这种方法还只是局限于用关键词来表征文档语义内容,因此文档之间的联系简化为关键词之间的关系,对文档的操作可转化为对关键词的操作,如:Mukherjea, Foley, and Hudson 所描述的系统能对文档的语义内容进行操作形成一个可视化层次结构[8]。在该系统中,文档根据其属性来组织,系统允许用户根据自己的信息需求指定文档属性,从而改变文档的显示结构。在 VR-VIBE 中,由用户确定关键词在金字塔中的位置,根据关键词在金定塔中的位置计算出每篇文档的距离从而确定文档在三维空间的位置[9]。LyberWorld's Relevance Sphere 也采用这种方法形成一个三维的文档集[10]。

此外,有的系统综合考虑文档的超链接和语义内容来进行可视化,如 Gershon et al. 所描述的系统允许用户察看通过链的层次结构所访问的文档结构图,也可以让用户根据自己的需求创建一个独立的层次图[11]。该系统还为被访问文档中的词建立一个同步图,帮助用户随时调整检索式中的词。

4 实例系统——TileBars

检索结果可视化的研究一直受到人们的关注,其中较为有代表性的研究成果包括:Scatter/ Gather, Tilebars, Tkinq 等。下面重点介绍 TileBars 系统。

4.1 TileBars 原理

TileBars 系统是由美国 Berkeley 大学数字图书馆研究人员开发出来的。该系统着眼于文档的内部结构,在检索结果的显示上,不仅提供了关于文档的信息,如文档的长度、文档标题,还揭示出检索词在文档中分布情况,如出现频率、出现在文档中哪个 page 等。这对于以往的以显示关键词和摘要为主的方式是一个突破。用户不仅能决定该看哪篇文档,还能决定看文档中具体哪一个 page 或哪几个 page。它在检索结果的提供上不再是笼统的整篇文档,而深入到文档内部,帮助用户快速找到最相关的内容。用户在输入检索式时,要把检索主题分成两组检索词(Term set 1 和 Term set 2),这两组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[12]。其检索结果文档的排序依据是:首先根据文档中出现了检索词的 page 数目的多少来排,然后根据出现检索词的数量来排,最后根据相似性检索来排。

其具体的显示方案为:将抽象的文档用矩形来表示,矩形的长度代表文档的长度,每个矩形中有上下两组方形,上面的方形代表命中第一组检索词的 page,下面的方形代表命中第二组检索词的 page,如图 2 所示。其中假定第一组检索词为 word 1,第二组检索词为 word 2。检索词与文档的相关度的大小用颜色的深浅表示,方形的颜色越深,说明检索词在该 page 中出现的频率越高(白色代表频率为 0,即检索词没有在该 page 中出现,黑色代表 8 或更多,频率是每组检索词中各个检索词的频率之和)。当文档中没有命中任何检索词的 page 数达到或超过 25 个时,用一个"X"来表示,矩形中所含的深色方形越多,说明该文档与检索词的相关度越高。用户需要浏览感兴趣的 page 时,只须点击具体的方形即可,而不必再为了查找具体的某一段而浏览整篇文档。

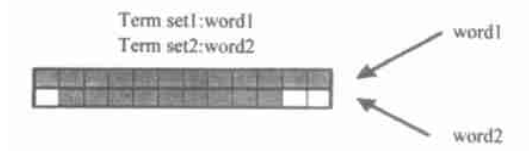

图 2　显示方案

4.2 评价

从对 TileBars 系统的分析来看,我们可以总结出该系统的成功之处在于它按照信息可视化的三要素完成了系统的构造。但从该系统将检索结果可视化所采用的方法来看,它没有揭示出文档簇内文档间的关系以及文档簇与文档簇之间的关系,不利于用户进行扩展浏览,因此查全率受到影响。

参考文献

1　华庆一,房鼎益. 三维可视化对于认知的作用. 计算机工程与科学,1998(3)
2　倪绍详. 可视化与 GIS. 地图,1996(1)
3　Gershon, N. D. 1994. (Panel chair.) Information visualization: The next frontier. ACM SIGGRAPH'94 Conference Proceedings, 485-486. New York: ACM
4　Robert Spence. Information Visualization. Addison Wesley. 2001.5

(下转第 53 页)

●谈春梅 段卫华 刘 伟

电子信息资源数据库检索系统的开发与实现*

摘 要 采用 XML 查询技术和显示格式,可开发出电子信息资源数据库检索系统,能实现模板基本查询、格式化显示及 XML 文档高级查询等功能。其乱码输出、模糊匹配等问题也可解决。参考文献 5。

关键词 电子信息资源数据库 XML 技术 网络检索 开发设计
分类号 G250.74

ABSTRACT By using XML query processing technology, we can develop systems for the access to electronic information resource databases, and actualize template basic search, formatted display, XLM file advanced search and other functions. The strange character output, fuzzy matching and other problems can also be solved. 5 refs.

KEY WORDS Electronic information resource database. XML technology. Network search. Development and design.
CLASS NUMBER G250.74

当前,电子信息资源组织的一种有效方式——电子信息资源数据库,已成为图书馆界开发研究的重要课题。我们借助于图书馆计算机网络平台,进行了电子信息资源数据库系统的开发设计。

电子信息资源数据库系统由 MARC 数据收集子系统、MARC 数据与大型关系型数据库相互转换子系统和电子信息资源数据库检索子系统 3 大模块组成。其中电子信息资源数据库检索子系统是电子信息资源数据库系统的重要组成部分。本文将着重介绍该子系统的主要开发技术及其程序的设计与实现。

1 系统的功能模块及主要开发技术

电子信息资源数据库检索子系统主要包括表单提问、网络基本查询、网络高级查询和查询结果数据显示等功能。

每个功能模块既相互独立又相辅相成。这些功

(上接第 50 页)

5 McCahill, M. P. & Erickson, T. Design for a 3D spatial user interface for Internet Gopher. Proceedings of ED-MEDIA 95, 39-44. Charlottesville, VA: AACE
6 Andrews, K. Visualising cyberspace: Information visualization in the Harmony Internet browser. Proceedings of Information Visualization, 97-104. 1995. Los Alamitos, CA. IEEE
7 Hendley, R. J., Drew, N. S., Wood, A. M., & Beale, R. Narcissus: Visualizing information. Proceedings of Information Visualization, 90-97. 1995. Los Alamitos, CA: IEEE
8 Mukherjea, S. Foley, J. D., & Hudson, S. Visualizing complex hypermedia networks through multiple hierarchical views. Proceedings of CHI'95, 331-337. New York: ACM
9 Benford, S., Snowden, S., Greenhalgh, C., Ingram, R., Knox, I., & Brown, C. VR-VIBE: A virtual environment for co-operative information retrieval. Eurographics'95, 349 - 360
10 Hemmje, M., Kunkel, C., & Willett, A. LyberWorld-A visualization user interface supporting fulltext retrieval. Proceedings of ACM SIGIR, 249 - 259. 1994. New York: ACM
11 Gershon, N. D., LeVesseur, J., Winstead, J., Croall, J., Pernick, A., & Ruh, W. Visualizing Internet resources. Proceedings of Information Visualization, 122-128. 1995. Los Alamitos, CA: IEEE
12 Marti A. Hearst. Tilebars: Visualization of term distribution information in full text information access. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, Denver, CO, May 1995. ACM

周 宁 武汉大学教授,博士生导师。通讯地址:武汉大学信息资源中心。邮编 430072。
文燕平 武汉大学信息管理学院 2001 级博士研究生。通讯地址:武汉。邮编 430072。 (来稿时间:2002-04-09)

*本文系江苏省哲学社会科学规划基金项目(省 05)研究成果。