●张学福

信息检索可视化基本问题研究*

摘　要　信息检索可视化的基本问题有: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。可根据实际情况合理选择,可选择一种,也可把多种技术组合在一起。表 1。参考文献 21。

关键词　信息检索　可视化映射　信息内容描述　可视化显示

分类号　G354

ABSTRACT　The author discusses the basic issues in the visualization of information retrieval: information retrieval models, information content representation, visualization mapping technology, visualization display technology, global mapping and local mapping, real-time visualization and human-intervened visualization, etc. 1 tab. 21 refs.

KEY WORDS　Information retrieval.　Visualization mapping.　Information content representation.　Visualization display.

CLASS NUMBER　G354

信息检索可视化是信息可视化技术在信息检索中的应用,是指把文献信息、用户提问、各类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[1]。信息检索可视化的基本问题主要包括: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、聚类分析、全局映射与局部映射以及实时可视化与人工参与的可视化等。

1　信息检索模型

信息检索可视化研究和实践应用首先要确定使用什么类型的信息检索模型。布尔逻辑模型、向量空间模型和概率模型是信息检索的三大经典模型。布尔逻辑模型的改进模型有模糊集合模型和扩展布尔模型;向量空间模型的改进模型有广义向量空间模型、潜在语义索引模型和神经网络模型;概率模型的改进模型有推理网络模型和信任度网络模型。

布尔逻辑模型通过对文献标识与提问式的逻辑比较来检索文献。向量空间模型把每一篇文献和每个提问均用向量表示,把信息检索中文献与提问的匹配处理过程转化为向量空间中文献向量与提问向量的相似度计算问题。概率模型是基于概率排序原理,根据文献与提问的相关概率来排序输出[2~4]。

信息检索可视化需要把文献描述为 N 维向量,然后在 N 维空间中显示出来,选用信息检索模型时要考虑哪种模型能够以向量的形式描述文献和提问。向量空间模型及其改进模型均符合这些条件。扩展布尔模型是先进行布尔检索,对检索结果按照向量检索模型处理,可视化信息检索模型可以根据实际情况采用扩展布尔模型、向量空间模型及其改进模型。

2　信息内容描述

文献信息内容描述常用的方法主要有:分类法描述,主题法描述,共频现象分析和概念图描述。其中前两种方法人们已十分熟悉。

共频现象分析是指如果两个词经常在同一篇文献中共同出现,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。共频现象分析包括:共引分析(作为引文共同出现),作者共引分析(作者在引文中共同出现)和词汇共频分析(两个专业词汇共同出现)等[5]。

概念图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。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。它能够促进用户吸收新概念和主题到他们已有的认知结构中。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[6]。

分类法利用由专家创建的等级结构的分类表标引文献,并且以先组式组配为主,灵活性较差,在可视化方面应用不多。主题法利用由领域专家创建的具有等级结构的,且词语规范的主题词表标引文献;共频现象分析使从信息资源中自动提取关联结构信息成为可能;概念图提供了一种基于用户的理解来描述语义结构的能力。这几种信息内容描述方法,可以单独使用,也可以根据需要,结合每种描述方法的特点,

* 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。

选择几种一起使用。

3 可视化映射技术

3.1 因素分析和主分量分析

因素分析(Factor Analysis)是一种多元研究方法,可以应用于分析一个大的数据集,可用于揭示变量之间关系的结构或对变量进行分类[7]。

主分量分析(Principal Component Analysis,PCA)是要素分析方法的核心方法,它可以把大量(可能)相关联的变量,转变成少量的不相关的变量(主分量)。第一个主分量说明数据尽可能多的变化性,每一个后继的主分量都说明余下数据尽可能多的变化性。相对于传统的聚类分析方法,因素分析的优势是它不要求每个对象必须归入一个簇中,可以以多种因素对对象进行分析。

3.2 多维测量

多维测量(Multidimensional Scaling, MDS),是试图在一组对象的相似测度中找到它们之间的结构[8]。它的实际作用是可以用来分析各种距离或者相似的矩阵。这些相似性可以表达人们对文献之间相似度、基于共频引文的对象之间的相似度等的评价。它的一个主要缺点是没有快速方法来解释降维后结果的自然特性。分析经常需要更多的局部细节和更多的明晰的结构表示。遇到这些需求时,MDS的配置会受到限制。另外,仅仅小型数据集可以用它来处理。

3.3 Kohonen's feature map (SOM)

Kohonen特征映射(SOM)是人工智能网络中主要的自组织学习方法之一。它把一系列高维数据映射到一个2维网格节点上,尽可能忠实地保持数据之间的关系[9~10]。

SOM算法采用了一组输入对象,每一个对象都用一个N维向量表示;把它们作为输入变量,并把它们映射到一个二维网格节点上。其结果是生成了一个有序的特征图。这种特征图有两个主要的特性。(1)尽可能忠实地保存输入数据之间的距离关系。映射保存了输入数据之间最重要的相邻关系,并使这些关系清楚显示出来。(2)特征图根据它们出现的频率,为输入向量分配不同数量的节点。以损害较低频率的样本为代价,频率较高的输入样本被映射到较大的区域。

SOM能将输入映射到低维空间,减少了维数,但高维映射到低维时会出现畸变,压缩比越大,畸变程度越大。SOM要求的输出神经元数很大,其权重向量数目也很多,这使它的神经网络规模较大。SOM在数据量较大情况下,随着学习次数的增多,学习效果反而降低,又称学习过度。

3.4 Pathfinder网(PFNET)

PFNET是由Fowler et al最早应用于可视化的。它根据经验性的数据,对不同概念或实体间联系的相似或差异程度作出评估,然后应用图论中的一些基本概念和原理生成一类特殊的网状模型。它对不同概念或实体间形成的语义网络进行表达,从一定程度上模拟了人脑的记忆模型和联想式思维方式,主要应用于认识心理学和人工智能等研究方面。在一般变换情况下,PFNET有一定的稳定性,并且通过对PFNET的分析,可以对不同的概念、实体进行分层和聚类。

PFNET清楚地显示了对象之间的链接,结构化的模式使人们观察非常方便;它是一种有效的减链接机制,保证一个网络不能含有太多的链接。若其节点数目很多,将会带来大量计算。随着节点数量的增多,会增加系统的负荷,降低系统的效率[11~12]。

3.5 潜在语义索引

潜在语义索引(Latent Semantic Indexing, LSI)的基本思想是文本中的词与词之间存在某种潜在的语义结构,因此采用统计方法寻找该语义结构,用语义结构来表示词和文本,达到消除词与词之间的相关性,简化文本向量的目的[13~14]。

潜在语义索引用正交的K维空间代替原来的空间,用该空间的点来表示词与文本,可能认为该空间是潜在的语义结构的概念空间,这样消除了词之间的相关性,降低了向量维数;在较低的概念空间,以更丰富的语义结构信息进行词与词和文本与文本的相关计算,如夹角余弦等。

它的优点:向量空间中每一维的含义发生了很大变化,反映的不再是词条的简单出现频度和分布关系,而是强化的语义关系;向量空间的维数大大降低,可以有效提高文本集的分类速度。

不足:降维因子K的选择带有很大主观性,K过大则运算量加大,K过小则会失去一些有用的信息。尽管用文档中包含的词来表示文档的语义,但其模型并不把文档中所有的词看做是文档概念的可靠表示。时间代价大。进行信息提取时,忽略词语的语法信息(甚至是忽略词语在句子中出现顺序),认为语法结构在文本的语义表达中处于次要的地位。

因素分析和多维测量方法虽然在一定程度上能够实现数据的降维,但不适用于数据集较大的情况。SOM方法,其更新程序不需要任何外部信号干涉,该算法是一个无监督的自组织算法。虽然在数据量较大的情况下,会出现学习过度的情况,但对从领域整体角度可视化领域知识的效果较好。PENET方法能够清楚地显示对象之间的链接,结构化的模式符合人们的认知模式;虽然要求节点数目不能太多,但对局部信息检索可视化来说,在某领域知识范围内能够较好地满足需求。LSI方法中包含矩阵的逆运算,信息的维数较少时,效果较好,超过40维就很难处理。在信息检索可视化具体应用时,要根据实际要求,选择合适的映射技术。

4 全局映射与局部映射

对具体数据信息进行映射计算,包括对大数据集映射计算和局部小数据集映射计算。全局映射与局部映射均以共频现象分析为基础。全局映射采用的是对领域整个数据集信息的大共频矩阵的映射策略,映射结果可反映整个领域的全貌。局部映射采用的是对领域数据集中局部信息的小共频矩阵的映射策略,虽然不能反映整个领域的全貌,但可以实现用户对映射过程的控制。

表1是White, et al对全局映射与局部映射的比较[15]。从表1可以看出,采用全局映射方式时,如果数据集确定,映射结果就是确定的,查询前映射图已经形成,用户不能改变它。采用局部映射方式时,不同的用户因需求不同,虽然对相同的数据集进行检索,但映射图是不同的,也就是说映射

图是动态生成的，有助于满足用户的特定需求。两种映射方法适合不同的需求，需要根据不同的需求进行选择。

表1 全局映射与局部映射的比较

全局映射	局部映射
设计者驱动	用户驱动
大矩阵映射	矩阵小子集的映射
查询前建立图	通过查询建立图
建立时间(单位)：小时	建立时间(单位)：秒
较难在图上标引	强调标引
通过在图上重定位进行探索	通过产生新图进行探索
用户是参观者	用户是指挥者

5 实时可视化与人工参与的可视化

5.1 人工参与的非实时可视化检索

这种检索是指在信息检索可视化过程中有专家参与选择关联词语(如作者)。可视化过程分成几个步骤。

这里以传统ACA的典型流程为例[16]。其流程为：选择作者；检索共引频率；编辑原始的共引矩阵；把共引矩阵转换为一个相关矩阵；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)。流程中包括许多人工处理过程和几个不同的用计算机处理的系统，例如用于文档检索的数据库，用于统计分析的统计分析软件(如SPSS)和用于可视化的图形处理系统等。由于统计映射过程，如多维测量和要素分析都包括大量的计算处理，很难与一个检索系统集成。

该种信息检索可视化形式的优点是词语选择经过领域专家参与，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。

5.2 动态实时可视化

动态实时可视化将信息检索可视化各个部分有机地结合为一体，不需要专家参与，由系统从数据库自动检索信息，完成数据资源选择、关联词语选择、建立关联词语共频次数原始矩阵、将原始矩阵转换为相关矩阵、用SOM、PFNETs等算法进行可视化映射、视图绘制等过程。它的优点是：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。动态实时可视化检索，是可视化信息检索的发展方向。

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。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，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。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。

用户对信息检索可视化系统是否要求实时可视化将会决定系统的复杂程度，具体设计系统时要考虑自己的具体需求和长远规划。

6 可视化显示技术

可视化显示技术主要用于把经过聚类处理的文献信息在计算机上以图形的形式显示出来。以下是主要的几种。

6.1 Focus + Context 技术

Focus + Context的思想最初源于Furnas关于鱼眼视图的研究[17~18]。它是一种放大显示画面中某块小的局部区域的透镜技术，放大区域的周围退到背景显示，但仍然可见。后来有很多人在这项技术的基础上创造了一系列新技术，允许用户在观察一个小的中心焦点区域的同时，保持一个较大周围区域的可见性，这就是Focus + Context技术的含义所在。这种技术可以将一个信息集合的特定部分的细节视图，通过某种方式和该信息集合的总体结构视图混合在一起；也可以认为是在显示一个大的信息空间的同时，其中的一部分以更细节的方式显示。

Focus + Context技术的理论根据：人们在现实世界中观察一个对象的时候，注重的是对象本身的细节，而对周围环境则不太在意；距离观察对象越远的周围环境引起人们注意的可能性越小；周围环境当然也有作用，它提示人们当前关注对象与周围环境的关系，同时也是转移注意的线索。Focus + Context以忽略细节的方式显示尽可能多的周围信息，将周围信息和以细节方式显示的焦点信息结合在一起，随着用户注意的变化改变细节显示部分。这种技术基于3个假设：用户同时需要概要信息(context)和细节信息(focus)；在概要和细节中需要的具体信息可能不同；这两种类型的信息可以结合在一个单一的(动态)显示中。而这正是人类视觉的观察特性。

Focus + Context技术的核心问题是保证focus信息正常显示的前提下，怎样才能显示更多的context信息，以及focus区域与context区域的方便切换。

6.2 Tree - map

Tree - map是Shneiderman等人提出的一种表示层次信息的可视化模型[19]。在这种结构中，层次结构的每个节点被表示为一个矩形，矩形的面积表示相应节点的权重。表示一个父节点的所有子节点的矩形被表示该父节点的矩形包围着。

Tree - map可视化结构充分利用了显示空间，通过一种

空间填充策略将层次结构映射为一个矩形。它使大型层次结构能够在有限的空间中显示,并且使语义信息的表示变得相对容易。Tree - map 将显示空间分割成互相包围的一些矩形,这些矩形表示树形结构。包围在某个矩形中的节点的画法完全依赖于节点的内容。每个节点的显示尺寸基于它相对于整棵树的比重。

Tree - map 可以有效利用计算机屏幕空间,并且能够很容易实现。但它丧失了层次结构的直观性,也丧失了对处于同一层次上不同父节点的子节点的关系(准兄弟关系),而这种关系对于把握节点之间的层次关系的结构特征是非常有用的。

6.3 Cone Tree

Cone Tree 是 Robertson, Mackinlay 和 Card 等提出的一种利用三维图形技术对层次结构进行可视化的方法[20]。其基本思想是利用三维图形技术将传统的二维树形表示法扩展到三维空间。在 Cone Tree 中,表示层次结构的整棵树以三维的形式进行组织和显示。利用三维锥形体来实现层次结构中父子节点之间的连接,层次结构的顶部放置在可视化空间的顶端,每个锥体的顶点表示该层结构的顶点,其子节点(三维)均匀排列在该锥体的底部。锥体的底面直径随着层次结构的深入逐渐减小,以保证最低层的结构也能在可视化空间中有效表示。每个锥体之间透明遮拦,可以保证每个锥体能够很容易被感知,还不会妨碍后面的锥体显示。同时辅以旋转、拖动等交互技术,可以很容易地实现对复杂层次关系的把握。

Cone Tree 可以比较容易地体现出树形结构的整体信息;在有限的屏幕空间中可以显示更多的节点;可以利用更多的手段来提供信息,如除了几何结构以外,还可以通过圆锥在平面上的阴影映射显示节点的分布情况。但空间分布极度不均匀,有的区域信息过于密集,而有的空间却未得到利用,尤其是在层次结构本身不平衡的情况下尤为突出;节点之间的相互遮挡情况严重。

6.4 Hyperbolic Tree

Hyperbolic Tree 是 Lamping 和 Rao 等提出的一种基于双曲几何的可视化和操纵大型层次结构的 Focus + Context 技术[21]。这种技术将更多的可视化空间用于显示层次结构中当前被关注的部分,同时又能把整个层次结构显示出来。它通过一种规范的算法将层次关系显示在一个双曲平面上,然后将这个双曲平面映射到显示区域。所选择的映射方式提供了一种鱼眼(fisheye)变形来支持 focus 和 context 之间的平滑过渡。

Hyperbolic Tree 把当前关注的节点移到显示区域的中心时,其父节点则被推移到周围,这时候很难分辨出父子关系;对处于同一层次上准兄弟关系,体现较难。

可以说,每种显示技术都有优缺点,在信息检索可视化时需要根据实际情况合理选择,既可以选择一种,也可以把多种技术组合在一起。

参考文献

1 赖茂生. 情报检索技术与方法的研究综述. 情报学进展(第五卷),2003
2 Ricardo Baeza Yates, Berthier Ribeiro Neto, et al. Modern Information Retrieval. ACM press, 1999
3 赖茂生等. 计算机情报检索. 北京:北京大学出版社,1993
4 苏新宁等. 信息检索理论与技术. 北京:科学技术文献出版社,2004
5,6 林夏. 信息可视化与内容描述. 现代图书情报技术,2004(10)
7,8 Katy Börner, Chaomei Chen, & Kevin Boyack. Visualizing Knowledge Domains. Annual Review of Information Science & Technology, Volume 37, 2003
9 Buzydlowski, jan William. A comparison of self-organizing maps and pathfinder networks for the mapping of co-cited authors. Drexel University, Ph. D. 2003
10 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
11 Schvaneveldt, R. W. (Ed.). (1990). Pathfinder associative networks: studies in knowledge organization. Norwood, NJ: Ablex.
12 Howard D. White. Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. Journal of the American Society for Information Science and Technology, 2003, 54(5)
13 Deerwester, S.; Dumais, S. T.; Landauer, T. K., Furnas, G. W, & Harshman, R. A. (1990). Indexing by Latent Semantic Analysis. Journal of the Society for Information Science, 41(6), 391 - 407
14 Chen, C., Czerwinski, M. 1998. From Latent Semantics to Spatial Hypertext: An Integrated Approach. Proceedings of the Ninth ACM Conference on Hypertext and Hypermmedia (Hypertext 98), ACM Press, pp. 77 - 86
15 White, H. D.; Lin, X.; Buzydlowski, J. W. Chen, C. (2004). User-Controlled Mapping of Significant Literatures. http://www.pnas.org/cgi/reprint/101/suppl 1/5297
16 McCain, K. W. Mapping authors in intellectual space: A technical overview. Journal of the American Society for Information Science, 1990, 41(6), 433 - 443
17 冯艺东. 信息可视化若干问题研究. 北京大学博士论文,2001
18,19,20,21 Stuart K. Card, Jock D. Mackinlay, Ben Shneiderman. Readings in information visualization: using vision to think. San Francisco, Calif.: Morgan Kaufmann Publishers, 1999

张学福 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,中国科学院文献情报中心、中国科学院研究生院博士生。通信地址:北京北四环西路33号中国科学院文献情报中心。邮编100080。

(来稿时间:2005 - 12 - 05)