软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(Suppl.(2)):186−192

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于相似性和距离中心性推荐的历史视图浏览*

谢江宁[1,2], 李学庆[1+], 唐 磊[1]

[1](山东大学 计算机科学与技术学院,山东 济南 250101)

[2](山东大学 研究生院,山东 济南 250100)

Historical Views Navigation Through the Recommendation Based on Similarity and Closeness Centrality

XIE Jiang-Ning[1,2], LI Xue-Qing[1+], TANG Lei[1]

[1](School of Computer Science and Technology, Shandong University, Ji'nan 250101, China)

[2](Graduate School, Shandong University, Ji'nan 250100, China)

+ Corresponding author: E-mail: xqli@sdu.edu.cn

Xie JN, Li XQ, Tang L. Historical views navigation through the recommendation based on similarity and closeness centrality. *Journal of Software*, 2012,23(Suppl.(2)):186−192 (in Chinese). http://www.jos.org.cn/1000-9825/12038.htm

Abstract: The visualization process usually only focuses on the results of the current visualization, losing the review and analysis of historical information. It implies that some important intermediate results are not tracked and detected timely, and often makes the information and the potential law invisible. The university graduate student information data is used as an example. When faculty in university visualize the students' personal information, they usually focus on the current view, ignoring the trace of the historical views, which contains important information or patterns. To address this problem and figure out the unknowns hidden in educational datasets, this paper proposes the historical views navigation through a similarity and closeness centrality based recommendation. In this approach, the useful intermediate views, or the interested views of the users are saved as history and compared with the current view. By analyzing the similarities between the closeness centrality and the path centrality, the most relative historical views are recommended to the user. Finally, the user study shows that most of the participants believe that this approach is a necessary alternative to properly integrate the current view and the historical views, which enhances user experience significantly.

Key words: history; visualization; recommendation; similarity; closeness centrality

摘 要: 用户在可视化过程中,通常仅仅关注当前的可视化结果,缺少对历史信息的回顾和分析,导致一些重要的中间结果没有被及时跟踪,不利于信息的对比和潜在规律的发现.以高校研究生信息数据为例,目前学校管理部门或学院对于研究生个人信息的可视化分析仅仅关注于当前数据,而忽略了更加重要的历史数据,从而无法准确地追踪和分析其潜在的信息和数据特征.针对这一问题,提出了基于推荐算法的可视化历史浏览方法.该方法将生成的可视化过程保存为历史信息,利用相似性算法分析历史结果间的关联关系,利用基于距离中心性和基于路径中心性的分析描述每个历史结果的重要性,综合 3 种分析方法将与当前可视化视图最为相似的历史信息推荐给用户,加速了认

* 收稿时间: 2012-05-15; 定稿时间: 2012-11-09

知过程.用户调查结果显示,大部分用户对基于历史的可视化推荐表示认可,并且已经开始在数据分析中使用.

关键词: 历史;可视化;推荐;相似性;距离中心性

信息可视化[1,2]是可视化技术的一个重要分支.该技术结合了数据挖掘、人机交互、认知科学、图形学等诸多学科的理论和方法,主要研究没有几何特征和空间信息的抽象信息的展现技术.经过多年的研究与发展,人们提出了多种可视化技术,较为常见的包括平行坐标轴、热度图、饼图、散点图等,这些可视化技术的出现为人们观察信息、理解信息提供了一种直观的方式,有助于人们认知过程的加速.

随着研究的深入,人们在可视化的基础上逐渐增加了分析和推理过程,多种分析算法被相继提出来,并被应用于可视化领域.基于相似性的分析是最常使用的方法之一.它利用欧氏距离、明氏距离、余弦值计算等方法计算数量间的相似程度,建立关系网络,使多维信息之间产生关联.基于图[3]的分析适合于具有网络结构的抽象数据,可以通过度、距离中心性、路径中心性、特征向量中心性度量方法分析每个网络节点的重要性.基于数理统计的分析方法主要以概率知识为基础,通过对大量事实进行统计分析,预测未来趋势,辅助决策的制定.另外,机器学习也是重要的分析方法之一.该方法通常用于信息的分类,通过对大量数据样本的学习,掌握信息和分类之间的映射关系,从而获取分类能力.

虽然人们提出了各种不同的可视化技术,用于信息的展现[4,5],并提出了多种方法用于数据的处理和分析,但这些方法都是以数据为中心的,仅仅关注最终结果的生成,缺少对可视化历史的分析.因此,当前的研究有两个不足之处:一是缺少对历史信息的可视化支持,对于一些重要的历史结果难以快速重现;二是不利于对比,同一组数据可能对应着不同的可视化技术和相应的视图,由于缺少对历史信息的保存和分析,无法实现相似结果的快速对比,不利于深入理解数据.针对可视化的研究现状和存在的上述问题,本文提出了基于推荐的可视化历史浏览方法.该解决方案将用户可视化过程中生成的中间结果和重要视图作为历史,以相似性分析、距离中心性和路径中心性度量作为分析依据,分析历史视图与当前用户生成视图之间的关联关系,将与当前可视化结果最为相似的历史信息推荐给用户,方便用户快速过滤、浏览相似结果,并实现不同可视化结果的对比.

1 相关工作

Dimsdale 和 Inselberg 设计并实现了平行坐标轴技术[6].该技术主要针对多维信息.多维信息中的每一维都由一条水平或者垂直的坐标轴表示.一条多维数据被绘制成通过所有坐标轴的折线,这条折线在每个坐标轴上的具体位置将根据其对应的实际取值进行映射.该方法的作用实际是将信息从高维空间映射到二维空间.

饼图是描述信息统计结果的一种重要方法[7].该方法最初由 William Playfair 在 1801 年提出,并被广泛用于商业信息和媒体信息.饼图通常被划分为多个扇形,每个扇形用于描述比重信息,所占空间越大,说明该扇形代表的信息量越大.

直方图是另外一种可视化技术,是一种统计报告图,通过聚类、数理统计等方法对信息进行概括和总结,由一系列高度不等的纵向条纹或线段表示数据分布的情况.一般用横轴表示数据类型,纵轴表示分布情况.该方法虽然原理简单,但是直观性较好,特别适用于不同信息属性间的比较.

欧氏距离是计算高维信息相似关系的重要方法[8].该方法通过比较高维空间中两个向量各个分量之间的距离计算相似性关系,两个高维向量距离越远,其真假性就越低.

基于图的分析方法通常会回答这样一个问题[3]:谁是网络中最重要的节点.根据对“重要”一词的不同理解,可以有多种分析方法,如基于度的分析、基于距离中心性的分析、基于路径的中心性分析和基于特征向量中心性的分析.每种方法从不同角度分析节点的重要程度.

本文在上述理论和方法的基础上提出了基于推荐的可视化历史浏览方法.该方法以直方图、饼图和平行坐标轴为基础,借助基于图的分析和相似性分析方法计算每个历史节点间的相似关系,以用户的当前视图为输入,推荐一系列相关的历史视图.

2 基于历史的推荐算法设计

首先定义历史的概念.本文的历史指的是来自同一数据集的多个可视化结果.这些历史结果可能采用了不同的可视化技术,也可能是同一数据集的不同子集.基于历史的推荐算法由两个子算法组成,分别对历史的相似性和每个历史的重要性进行分析.它以用户当前浏览的视图为基础,向其推荐最为相关的多个历史.

2.1 相似性分析

每一个可视化历史结果都对应着一组多维数据,因此,历史视图之间的相似性关系可以描述为对应数据集之间的相似关系.设历史视图 i 对应的 n 维数据集为共包含 m 条数据,第 k 条数据可以表示用 $P_i^k=(P_i^k(1), P_i^k(2),...,P_i^k(n)),1\leqslant k\leqslant n$ 表示,则整个数据集对应视图的平均向量可以表示为

$$P_i=\frac{\sum_{k=1}^{m}P_k}{m}.$$

向量代表了视图 i 对应数据集所有数据的平均值.它是一个 n 维信息,代表了数据集的 n 个属性.利用平均向量可以计算每个历史视图之间的相似关系.设当前的可视化视图为 c,历史视图对应 q,则每个历史图与当前视图的相似度 $S(c,q)$可以通过如下几种方法计算求得:

四则运算法(arithmetic):

$$S(c,q)=\sum_{k=1}^{n}\frac{1-\frac{|P_c[k]-P_q[k]|}{P_c[k]+P_q[k]}}{n}.$$

几何运算法(geometric):

$$S(c,q)=\sqrt[n]{\prod_{k=1}^{n}\left(1-\frac{|F_c[k]-F_q[k]|}{F_c[k]+F_q[k]}\right)}.$$

非开方几何运算法(geometric no root):

$$S(c,q)=\prod_{k=1}^{n}\left(1-\frac{|F_c[k]-F_q[k]|}{F_c[k]+F_q[k]}\right).$$

欧氏距离法(Euclidean):

$$S(c,q)=sqrt\left(\sum_{k=1}^{n}(F_c[k]-F_q[k])^2\right).$$

经过实验发现,欧氏距离法的计算效率最高,因此本文采用此方法作为视图间相似计算的主要依据.为了实现归一化计算,本文实际使用的是反转欧氏距离公式:

$$S(c,q)=\frac{1}{1+sqrt\left(\sum_{k=1}^{n}(F_c[k]-F_q[k])^2\right)},\ n=73,\ 1\leqslant q\leqslant m,\ q\neq c.$$

通过欧氏距离,可以求得每个历史图与当前视图之间的相似值.

2.2 基于距离中心性的历史分析

首先建立历史信息网络图,利用上节中的相似性计算方法计算所有视图(包括历史视图和当前视图)之间的相似性关系,根据计算结果生成相应的网络图.其中,每个节点代表一个可视化视图,节点之间的边表示两个可视化结果之间的相似性关系,边的权值越大,相似性越高.利用历史信息网络图和基于图的分析方法,可以回答这样一个问题:谁是网络中最重要的节点.

基于图的分析有多种方法.基于度的方法主要考虑每个节点连接的其他节点的数量,一个节点连接的其他节点越多,其重要性越大.基于距离中心性的方法主要考虑某个节点与其他节点间的平均最短距离[9].该方法通常用于最小耗费的计算.基于路径中心性的方法考虑某个节点在每对节点间最短路径中出现的次数[10].基于特征向量中心性的方法同时考虑了节点的度和与之连接的边的权值[11].其中,基于度的方法过于简单,基于路径的

方法在描述重要性方面不够准确,基于特征向量中心性的方法虽然考虑全面,但由于前文已经计算了相似性关系,已经不需要考虑权值问题,因此,使用基于距离中心性的方法在本文中较为合适.

在计算某个节点的距离中心性时,首先需要求出这个节点与网络中其他每个节点之间的最短距离,也就是最短路径[12],将这些最短路径长度相加取平均值就得到了该节点的距离中心性.它表示该节点与网络中其他节点之间的平均最短距离.一般来说,一个节点的距离中心性越小,表示它与其他节点之间的关系越近,与其他节点的通信成本就越低,在网络中的重要性也就越大.这样的节点通常在整个网络中居于中心地位.当然,也有部分学者认为,节点的距离中心性的取值越大,它在网络中的地位就越高,也就越重要.这取决于使用环境和实际需要.一个需要注意的问题是,一个节点的距离中心性需要计算该节点与其他节点之间的最短路径,但并不是每对节点之间都是可到达的,可能它们之间并不存在最短路径.对于这样的情况,一个通常的解决方法是忽略不可达路径,只计算一个节点与其他可达节点之间的平均最短路径.距离中心性的形式化描述如下:

$$c(v_i)=\frac{\sum_{j=1}^{m}d(v_i,v_j)}{n},\ j\neq i.$$

其中,$c(v_i)$ 表示节点 v_i 的距离中心性的值,m 表示网络中节点的数量,$d(v_i,v_j)$ 表示节点 v_i 和 v_j 之间的最短路径,但不计算与自身的最短路径和不可达的路径,n 表示所有可达的最短路径的数量.利用上述公式计算网络中所有节点的距离中心性,结果越小,表示这个节点与其他节点的关系越近,在网络中发挥的作用也越大,在网络中倾向于中心位置.

利用上述两种分析方法,可以最终求得每一个历史视图与当前视图之间的关联关系.当用户对于信息中的某些数据类型产生兴趣时,系统会以此为依据通过相似性和距离中心性对相关信息数据进行分析,产生与该数据相关的历史列表.同时对该列表进行排序并且形成推荐.

2.3 基于路径中心性的历史分析

路径中心性是衡量一个节点重要性的另一种方法.一个节点的路径中心性与距离中心性代表的意义不同.距离中心性表示的是距离关系,而路径中心性表示的是依赖关系,它表示网络中其他节点对某一节点的依赖程度.如果网络中的大部分节点都对某一节点产生依赖,就说明该节点在整个网络中的影响力比较大,重要性高;如果缺失该节点,则整个网络的结构和节点的相互关联关系会受到较大影响.要体现网络中多个节点对某一节点的依赖程度,可以用“经过”来描述.要计算网络节点 v 的路径中心性,首先需要求出网络中任意一对节点之间的最短路径,然后计算这些最短路径之中,经过节点 v 的路径数量,包含节点 v 的最短路径的数量在所有路径中所占的比例就是节点 v 的路径中心性.

对于本文的研究内容,要实现基于路径中心性的历史分析同样需要一张网络图,其生成方法见上一节,即节点代表一个可视化结果,边代表每个视图之间的关联关系.接下来,计算网络中每个节点在网络中的路径中心性.首先计算每对节点之间的最短路径,然后计算每个节点在所有路径中出现的次数,次数越多,说明该节点越重要,一旦失去该节点,就会产生大面积的网络瘫痪.其公式化描述如下:

$$b(v)=\frac{\sum_{s\neq v\neq t}d_{st}(v)}{\sum_{s\neq v\neq t}d_{st}}.$$

其中,$b(v)$ 代表网络中节点 v 的路径中心性,s 和 t 表示不同于节点 v 的节点,$d_{st}(v)$ 表示节点 s 和 t 之间,经过节点 v 的最短路径,$d_{st}(v)$ 有两个取值,当 s 和 t 之间的最短路径包含 v 时,$d_{st}(v)=1$. 当 s 和 t 之间不存在最短路径或者该路径不包含节点 v 时,$d_{st}(v)=0$.

2.4 算法融合

本文从 3 个方面对视图间的关系进行了分析,包括相似性分析、基于距离中心性的分析和基于路径中心性的分析.这 3 种算法最终以一定的方式进行融合,对视图间关系进行综合分析.设当前用户感兴趣的节点为 v,节点 p 为其他节点,$S(v,p)$为两个节点间的相似性,$C(v,p)$表示基于距离中心性的结果,$B(p)$表示节点 p 的路径中心性,则节点 p 及其关联程度 $R(v,p)$可以通过如下公式生成:

$$R(v,p)=a\times S(v,p)+b\times C(v,p)+(1-a-b)\times B(p),\ 0\leqslant a,\ b\leqslant 1,\ 0\leqslant a+b\leqslant 1.$$

其中,a,b 表示混合因子,代表了这 3 种算法在对于最终结果的影响,a 越大,相似性分析结果所占比重越大;b 越大,基于距离中心性的算法所占比重越大.

3 用户接口和交互设计

针对高校研究生信息数据的特点和管理人员的分析习惯,本文设计了平行坐标轴、饼图和直方图 3 种可视化技术[13].可视化界面如图 1 所示.整个分析系统的布局分为 3 个部分,功能区、可视化主界面和历史推荐列表.功能区集成了一些常用的功能按钮,如导入、过滤数据集,加载历史、选择可视化技术等.可视化界面用于显示结果.历史推荐显示了与当前视图最为相关的历史信息.这些信息由上一节中的推荐算法生成,用户可以通过方向按钮查看其他历史内容.

图 1 用户接口

当用户过滤、加载相关信息数据并指定一种可视化技术以后,相应的结果会在主视图中显示.并且相应的数据集会以参数的形式传递给推荐算法.推荐算法以当前的可视化结果为依据,计算每个历史结果的相似度,按照相似度的高低对整个历史列表进行排序,将排名靠前的、与当前视图最相关的结果展现给用户.其结果如图 2 所示.

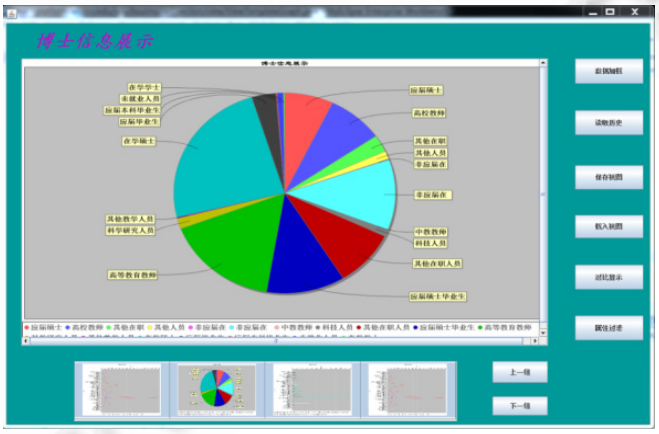

图 2 当前视图与推荐结果

根据用户针对可视化分析的需求,本文还设计了数据的对比功能,方便用户将历史视图与当前视图进行对

比.当用户对推荐算法给出的某个结果感兴趣时,通过数据对比功能实现 side-by-side 对比,可以观察相同数据集的不同可视化结果,也可以查看同一数据集的不同子集之间的差别.图 3 是两种不同可视化技术的对比结果,从不同侧面描述同一数据集的特征.通过对比,用户可以清晰地了解和选择适合自己数据类型和分析需求的可视化技术.

图 3 结果对比

4 实 验

本文以用户调查的形式对推荐算法和可视化、交互技术进行了验证,将基于历史的可视化推荐用于山东大学研究生管理信息系统.本文设计了 3 种测试方式,分别是:正常模式,仅包含可视化和交互功能;基于历史的可视化模式,包含 3 种可视化技术和用户浏览历史;基于历史的可视化推荐,即本文研究内容.整个测试涉及研究院及选取的相关学院管理类人员 30 人,测试过程持续 2 周,共 10 个工作日.测试以量化打分的形式进行,被测人员每天对上述 3 种模式进行测试,并对每种模式进行打分,分数范围为 0~10.经过对最终的反馈信息进行统计,其结果如图 4 所示.其中,mode 1 由于缺少历史信息的支持,整体结果呈水平变化趋势,得分最低.mode 2 增加了用户历史的浏览,用户使用初期,用户历史积累较少.该模式得分呈上升趋势,但随着可视化历史的增加,用户难以从中快速挑选重要的视图,导致得分开始下降.该模式整体得分高于 mode 1.本文设计的基于历史的可视化推荐在开始时得分与 mode 2 相同,但随着用户历史的增加,推荐结果的匹配度会越来越好,用户满意度呈递增趋势,整体得分最高.

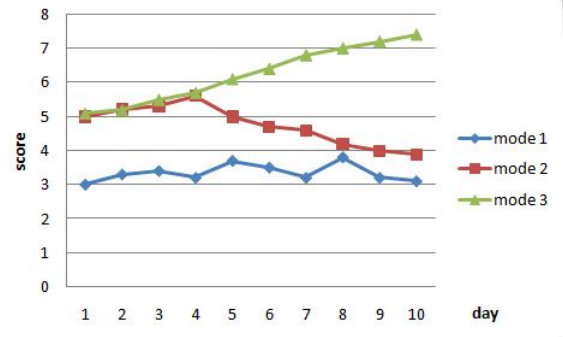

图 4 测试结果对比,其中 mode 1~ mode 3 分别对应正常模式、基于历史的模式和基于历史的可视化推荐

5 总 结

本文提出了一种基于推荐的可视化历史浏览方法.该方法以可视化历史和重要中间结果为主要依据,采用

了两种分析方法对可视化历史进行分析,一种是利用欧氏距离分析历史视图与当前视图间的数据相似性关系,另一种是利用基于图的分析方法分析历史视图相对于当前视图的重要性.综合两种分析方法,最终生成推荐结果,将与当前视图最为相关的结果推荐给用户.用户测试结果表示,大部分用户对本文研究内容感兴趣,对可视化界面和交互技术的认可度较高,对推荐结果也较为满意.大部分用户表示愿意继续使用基于历史的可视化推荐.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是山东大学研究生院薛佩军主任、计算机科学与技术学院李学庆教授和唐磊表示感谢.

References:

[1] Plaisant C. The challenge of information visualization evaluation. In: Proc. of the Working Conf. on Advanced Visual Interfaces. 2004. 109−116.

[2] Keim DA. Information visualization and visual data mining. IEEE Trans. on Visualization and Computer Graphics, 2002,7(2): 100−107.

[3] Newman MEJ. The Mathematics of Networks. 2nd ed., Basingstoke: Palgrave Macmillan, 2008.

[4] McDonnell KT, Mueller K. Illustrative parallel coordinates. Computer Graphics Forum, 2008,27(3):1031−1038.

[5] Bachthaler S, Weiskopf D. Continuous scatterplots. IEEE Trans. on Visualization and Computer Graphics, 2008,14(6):1428−1435.

[6] Inselberg A, Dimsdale B. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In: Proc. of the 1st Conf. on Visualization. 1990. 361−378.

[7] Franklin KM, Roberts JC. Pie chart sonification. In: Proc. of the Information Visualization (IV 2003). 2003. 4−9.

[8] Danielsson PE. Euclidean distance mapping. Computer Graphics and Image Processing, 1980,14(3):227−248.

[9] Okamoto K, Chen W, Li XY. Ranking of closeness centrality for large-scale social networks. In: Proc. of the 2nd Annual Int'l Workshop on Frontiers in Algorithmics (FAW 2008). 2008. 186−195.

[10] Vuki D. Network descriptors based on betweenness centrality and transmission and their extremal values. Les Cahiers du GERAD, 2011.

[11] Carreras I, Miorandi D, Canright GS, Engo-Monsen K. Eigenvector centrality in highly partitioned mobile networks: Principles and applications. Studies in Computational Intelligence, 2007,69:123−145.

[12] Borgwardt KM, Kriegel HP. Shortest-Path kernels on graphs. In: Proc. of the 5th IEEE Int'l Conf. on Data Mining. 2005. 74−81.

[13] Dix A. Human-Computer interaction: A stable discipline, a nascent science, and the growth of the long tail. Interacting with Computers, 2010,22(1):13−27.

谢江宁(1980－),男,山东济宁人,博士生,主要研究领域为信息数据可视化分析.

唐磊(1981－),男,博士,主要研究领域为人机交互,可视化分析.

李学庆(1964－),男,博士,教授,博士生导师,主要研究领域为数字化设计,软件工程,人机交互,可视化.