面向拓扑感知的层次结构信息可视探索方法

谭博友[1],韩永国[1],王桂娟[1],赵韦鑫[1],周 锐[1],蔡梦杰[1],吴亚东[2]
(1. 西南科技大学 计算机科学与技术学院,四川 绵阳 621000;
2. 四川轻化工大学 计算科学与工程学院,四川 自贡 643000)

摘 要:在有限的屏幕范围内,用户从有分支拥挤和节点遮蔽的层次可视化视图中获取拓扑结构信息具有挑战性。针对以上难点,提出了一种面向拓扑感知的层次结构信息探索框架。为提高用户探索拓扑结构信息的效率,提出采用重要节点评估算法。通过对以重要节点为根的子结构以视觉编码的形式进行隐藏,同时确保在保留较多的结构信息的条件下解决了分支拥挤节点遮蔽等问题。基于文本关键词的思想,定义了一种层次数据关键子结构的提取方法,通过提取关键子结构对整体拓扑结构信息进行概要,帮助用户理解整体拓扑结构特征。为提高用户对相似子结构的探索对比分析的效率,基于图表示学习算法将层次结构的节点进行向量化表示,通过将节点向量进行高斯混合聚类来构建相似子结构集合,然后采用图核计算子结构的相似度分数,通过相似度分数排序后完成相似子结构的提取。基于以上算法,设计了一个交互式的可视分析系统。通过可视分析系统完成了两项案例分析和两项用户实验,证明了所提框架的有效性。

关键词:层次结构数据;重要节点评估;关键子结构提取;层次结构向量化;可视分析
中图分类号:TP391.41 **文献标识码**:A **文章编号**:1673-629X(2022)11-0081-07
doi:10.3969/j.issn.1673-629X.2022.11.012

Visual Exploration Method of Hierarchical Structure Information for Topology Awareness

TAN Bo-you[1],HAN Yong-guo[1],WANG Gui-juan[1],ZHAO Wei-xin[1],
ZHOU Rui[1],CAI Meng-jie[1],WU Ya-dong[2]
(1. School of Computer Science & Technology,Southwest University of Science & Technology,
Mianyang 621000,China;
2. School of Computer Science & Engineering,Sichuan University of Science and Engineering,
Zigong 643000,China)

Abstract:In a limited screen range,it is challenging for users to obtain topology information from hierarchical visualization views with branch congestion and node shadowing. In view of the above difficulties,a hierarchical information exploration framework for topology awareness is proposed. In order to improve the efficiency of users in exploring topology information,an important node evaluation algorithm is proposed. The sub-structure rooted in important nodes is hidden in the form of visual coding,while ensuring that the problems such as branch crowded node masking are solved under the condition of retaining more structural information. Based on the idea of text keywords,a method for extracting the key substructure of hierarchical data is defined,which summarizes the overall topology information by extracting the key substructure to help users understand the characteristics of the overall topology. In order to improve the efficiency of users' exploration and comparative analysis of similar substructure,the nodes of hierarchical structure are represented by vectorization based on graph representation learning algorithm,and the set of similar substructure is constructed by Gaussian mixture clustering of node vectors. Then the similarity score of the substructure is calculated by using the graph kernel,and the similarity substructure is extracted after the similarity score is sorted. Based on the above algorithms,an interactive visual analysis system is designed. Two case studies and two user experiments are completed through the visual analysis system to prove the effectiveness of the proposed framework.

Key words:hierarchical structure data;important node evaluation;key substructure extraction;hierarchical structure vectorization;visual analysis

收稿日期:2021-11-19 **修回日期**:2022-03-22
基金项目:国家自然科学基金(61872304,61802320)
作者简介:谭博友(1996-),男,硕士研究生,CCF 会员(88054G),研究方向为层次数据可视化;通讯作者:吴亚东(1979-),男,博士,教授,博导,CCF 杰出会员(13866D),研究方向为可视化、人机交互、虚拟现实。

0 引 言

层次结构数据不仅常见于文件系统、动植物分类、系谱图等领域,还能将复杂领域通过层次结构的形式进行简化表示[1]。层次数据的拓扑结构,能直观地反映某一事物的分类特征,帮助数据分析人员探寻结构信息并理解层次数据的分类模式等。因此,探索层次数据的拓扑结构信息是层次数据分析的一项重要环节。

通过可视化的方式探索层次数据比较常见。现有的层次数据可视化技术主要分为连接式、空间填充和混合式[2]。层次拓扑结构的绘制技术通常基于 Reingold 和 Tilford's[3]算法,利用模块化方法来定位节点,其中子节点位于父节点下方(自上而下的方向),或者位于右侧(从左向右的方向)。由于显示空间的限制,原始的经典布局主要在一维上扩展,降低了层次结构数据可视探索的实用性,同时伴随节点数的增多也带来了节点遮蔽和分支拥挤等情况。为了克服上述限制,各种交互技术应运而生,例如缩放、鱼眼视图和一维扭曲。流行的交互式可视化方法则通过显示感兴趣的子结构同时缩小其他子结构来提供焦点+上下文视图。如 SpaceTree 和 DOITree。此外,视觉线索提供有关隐藏分支内容的持续信息,并以平滑的动画显示节点何时被聚焦或缩小。对隐藏分支使用视觉编码提示在探索大型层次结构方面已经证明是有效的[4-5]。但在寻找目标子结构或相似子结构时,需频繁地点击视觉提示展开隐藏的分支,这将浪费用户大量的时间,同时对用户的瞬时记忆也带来了挑战。对于采用兴趣度进行分支隐藏的方法,需要用户指定各节点的初始兴趣度,这对于用户来说具有一定的挑战性。若节点未指定兴趣度,如何确定那些分支应该隐藏并确保未隐藏的分支能为用户呈现更多的结构信息,降低用户在探索的过程中频繁点击的隐藏分支的时间消耗也具有挑战性。

基于以上问题,该文提出了一个面向拓扑感知的层次结构信息探索框架。该框架通过评估重要节点来解决无兴趣度标记时分支隐藏的问题并保留更多的结构信息。通过提取关键子结构对整体层次结构进行概要,同时引入图嵌入算法对结构进行向量化来提高用户探索子结构和相似子结构的效率。最后,基于该框架实现了一个交互式的可视探索系统,通过案例分析和用户实验来验证该方法的有效性。

1 层次拓扑结构信息计算

层次结构信息计算主要由三大模块组成,分别为重要节点评估模块、关键子结构提取模块、相似子结构提取模块。

1.1 重要节点评估

为解决在无兴趣度标记的层次结构数据分支隐藏的问题,该文引入网络分析领域中提取重要节点的分析方法,对层次数据中的重要节点进行提取。在对网络中的节点进行重要节点评估的算法中,常采用度指标、接近度指标和介数指标。其中接近度指标和介数指标都需要计算各节点之间的最短路径,但层次结构数据大多都只有一个根节点,大多路径都将通过根节点,这将影响算法的计算结果,因此接近度指标和介数指标不适用层次结构的节点重要性评估。度指标则采用计算节点的度值来确定节点的重要性。除度指标以外,张玫等[6]提出的合度评估算法以及 DN(delete node based on both degree)评估算法均围绕节点的度值进行评估,为确定最佳的评估算法,依次对上述算法进行实验,结果如图 1 所示。从图中可以看出,在保证没有节点遮蔽以及分支拥挤的情况下,DN 算法的结果保留了更多的节点、分支结构和深度信息。

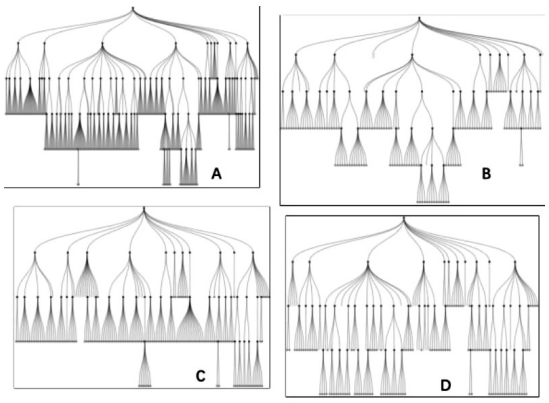

图 1 重要节点评估算法结果对比
(以食物分类数据为例)

A:原始布局 B:度指标评估算法结果 C:合度算法结果 D:DN 算法结果

通过以上实验结果,该文采用 DN 评估算法来对层次数据的节点的重要性进行评估。由于不同数据的节点数量不一样,其所要收缩隐藏的节点数是不确定的,因此,在实际应用中基于 DN 算法进行了扩展。以此适应不同的节点数量的层次数据。其算法主要流程为:

Step1:根据 DN 算法计算各节点重要度;

Step2:根据重要度排序,取 Top-k 节点,k 默认取 10;

Step3:将这些重要节点及其子节点进行隐藏,保存这些重要节点;

Step4:通过布局算法计算各节点位置;

Step5:当前兄弟节点之间圆心距离减去一个直径后是否小于阈值 n,是则将当前绘制出的所有节点重复 Step1 ~ Step4;

Step6:输出重要节点序列。

流程中的布局算法主要指基于节点连接方法的正交布局和径向布局两种,空间填充和混合式涉及的布局算法并不适用于该文。文中食物分类数据为实验数据,通过调节阈值 n 来对比两种算法对于重要节点数量的影响。实验结果显示,在相同的阈值 n 下径向布局算法提取的重要节点数量少于正交布局,且伴随 n 的增长,径向布局的效果越明显。采用正交布局时,其对于层次结构的分层效果展示更加直观,径向布局的直观效果没有正交布局那么好。因此,如果考虑隐藏更少的重要节点,建议采用径向布局;如果追求更好的层次分明可视化效果,建议采用正交布局。

1.2　关键子结构提取

Ben Shneiderman 曾提出“Overview first, zoom and filter, then detail on demand”的标准分析流程[7]。该流程指出在可视分析中概览属于第一步。因此,要更好地对层次数据的拓扑结构进行探索分析,需提供关于结构的概览。基于 TF-IDF 算法提取文本关键词的思想,该文通过提取关键子结构来高度概括整体层次结构。

在提取关键子结构之前,首先要解决子结构的定义问题,定义一种类似于文本中的词的子结构,这样才能更好地提取关键子结构。

在层次结构中,其最小单元为节点,但一个节点并不能算是一个拓扑结构,一个拓扑结构至少应包含边。因此,定义的子结构应由节点和边组成。在层次结构中,由边和节点组成的最小单元为父子节点连接而成的结构。因此,为便于对整个层次结构进行分解,对子结构定义如下:非叶子节点和其直接相连的节点构成一个子结构(如图 2 中 A 区域所示)。

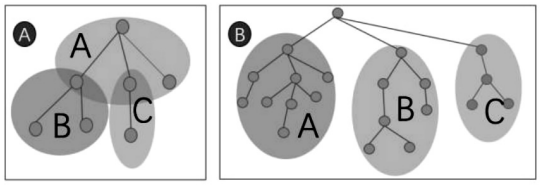

图 2　A 区域:子结构示意图

(A、B、C 区域代表三个子结构)和 B 区域:层次结构文档的划分示意图(A、B、C 代表三个结构文档)

由于 TF-IDF 算法由词频 TF 和逆文档频率指数 IDF 两部分组成,因此要提取关键子结构,首先需计算 TF 和 IDF 值。计算 TF 值相当于计算子结构的频率。由于文本中计算 IDF 时通常需要文档库,因此计算层次结构的 IDF 时也需要类似于文档库的东西。但是层次结构数据往往是一个整体,如果直接将整体作为一个文档计算 IDF 将得到 0 值。因此,在计算 IDF 值时应定义层次结构的文档库。

该文采用以下方法定义层次结构的文档库。首先,假设根节点直接相连的节点有多个,这时删除根节点,可得到多个如图 2 中 B 区域 A、B、C 所示的子结构。将图 2B 区域 A、B、C 这样的子结构单独作为一个层次结构数据,那么 A、B、C 就可以构成当前层次结构的文档库。

根据以上定义,列出如下 TF 和 IDF 计算公式:

$$\mathrm{tf}_{ij} = \frac{C_{ij}}{\sum_k C_{kj}} \tag{1}$$

$$\mathrm{idf}_i = \log \frac{|T|}{|\{j: t_i \in S_j\}|} \tag{2}$$

公式(1)中,C_{ij} 表示子结构 i 在结构文档 j 中出现的次数,分母表示结构文档 j 中 k 个子结构出现次数和。公式(2)中,T 表示结构文档总数,分母表示 j 结构文档中子结构 i 出现在所有结构文档中的次数。tf_{ij} 与 idf_i 的乘积代表 j 结构文档中 i 子结构的 tf-idf 值。然后将各结构文档中的各子结构的 tf-idf 值进行降序排序,取前 n 个作为该结构文档中的关键子结构。最后将所有结构文档的关键子结构进行汇总去重后,剩余的子结构作为整个层次结构的关键子结构。

1.3　相似子结构提取

相似子结构是对整体结构的一种过滤,能帮助用户快速定位与用户焦点子结构相似的结构特征。

相似子结构的匹配实际为对相似子图挖掘。现有的许多工作总结了近年来的子图挖掘相关算法,如 VF2plus[8]、VF3[9]等。这些算法虽然能够得到完全匹配的子图,但在匹配的过程中存在会匹配大量不符合的结构导致大量的运算时间消耗。

笔者希望用户在使用该可视探索系统时能在较短的交互时间内获取结果,若采用 VF3 等算法,将导致在探索的过程中浪费用户大量的时间而影响交互体验。潘嘉铖[10]和李珍[11]等人的研究表明,通过表示学习将网络中的节点进行向量化,能减少无关的子结构的匹配次数,极大地提高相似子结构的提取效率,同时生成的节点向量更能体现“邻居紧密性”。因此,该文采用表示学习的方法对层次结构的节点进行向量化。

采用 GraphWave[12]算法对层次结构进行向量化。所需提取的相似子结构存在于整个结构的不同位置,只需局部结构特征具有相似性即可。GraphWave 算法已被证明能很好地提取图中局部特征相似的结构。

为保证提取出来的相似子结构的准确性,采取以下步骤提取相似子结构:

Step1:对向量化后的节点使用高斯混合聚类模型进行聚类,目的是将相似的节点聚集到同一类簇,在相同聚类簇中的节点所构成的子结构的相似性也更

接近。

Step2:将用户当前所选子结构的节点向量所属的所有聚类簇中的所有节点作为相似子结构匹配候选集。

Step3:剔除用户所选子结构上的节点,在剩余的节点集中构建子结构集合。由于构建出的子结构存在其节点数量远小于或者远大于用户所选子结构节点数量的问题,因此将该部分结构删除,得到候选的相似子结构集合。

Step4:采用 Weisfeiler-Lehman 图核[13]对候选相似子结构集合中的子结构分别与用户所选的子结构进行相似度分数计算,相似度分数越高的子结构其相似度越高。按相似度分数降序排序,取前 k(默认 $k=6$)个相似子结构呈现在相似子结构查看面板中。

为证明文中相似子结构提取算法的效率,与 VF3 算法在食物分类数据(总共 463 个节点)上进行了对比,通过测试的数据显示,VF3 算法的平均时间在 15 到 20 秒之间,而文中算法平均在 5 秒内就能返回结果。因此,采用图表示学习将节点向量化后,对相似子结构的效率是有所提升的。

2 可视化系统设计

为便于用户交互式地探索层次数据的拓扑结构,基于层次结构探索框架开发了原型系统,系统界面如图 3 所示。系统主要分为三个模块:(1)数据概览模块(A1、A2、A3);(2)主视图模块(B1、B3);(3)相似子结构探索模块(C1、C2、C3、C4)。

2.1 数据概览模块

数据概览模块主要目的是为用户提供层次数据的概览。其主要由向量聚类投影视图(A1)、节点统计概览视图(A2)、度分布概览视图(A3)组成。

由于 GraphWave 算法生成的节点向量为高维向量,难以直接可视化呈现。为便于用户更直观地探索相似子结构和对节点的整体特征进行概览,首先使用高斯混合模型将节点向量聚类得到各节点的聚类标签,然后通过 T-SNE 将各节点高维向量投影到二维平面,最后基于聚类标签对投影平面中的各节点进行颜色编码。在向量聚类投影视图中用户可通过点击不同颜色的点,以查看相同颜色编码的节点在节点连接视图中的分布情况。

由于主视图只是对层次结构的呈现,对于如节点度值分布等信息无法很好地体现。因此,设计了两个统计性的数据概览视图帮助用户更好地了解数据的拓扑结构信息。以堆叠柱状图(A2)的方式显示了当前系统所探索的层次结构数据在总体上的层数、每层节点的数量、叶节点和非叶子节点在各层之间的占比。同时,通过气泡图(A3)的形式来呈现每层节点的度值统计情况,气泡的大小为度值相同的节点数量。这能帮助用户直观地了解当前探索的层次数据各层节点孩子节点数的分布情况。

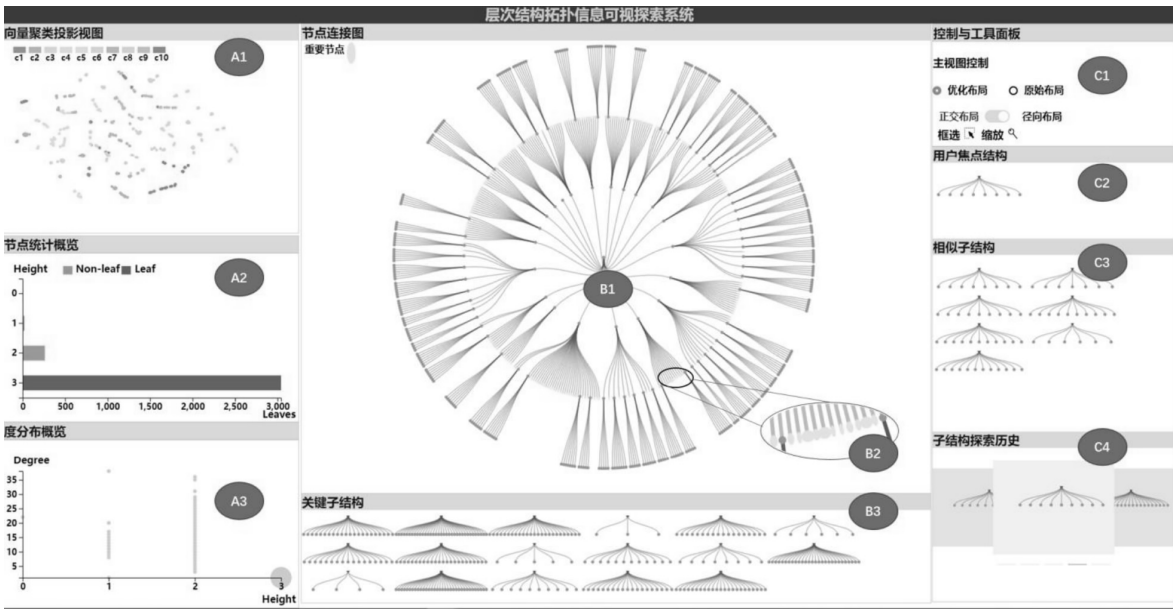

图 3 层次结构拓扑信息可视探索系统

A1:向量聚类投影视图 A2:节点统计概览视图 A3:节点度分布概览视图 B1:节点连接视图 B2:重要节点视觉编码形式 B3:关键子结构概览视图 C1:控制与工具模块 C2:用户焦点拓扑结构 C3:top-k 相似拓扑结构 C4:子结构探索历史

2.2 主视图模块

主视图中的节点连接视图(B1)为用户探索层次结构数据结构信息的主要视图,在该视图中,结合了节点重要性评估算法的评估结果,将对布局质量影响较大的节点进行隐藏。在其父节点处进行视觉编码(B2)以提示用户该节点处存在隐藏的节点及其子结

构。椭圆的短轴表示其隐藏的子结构的深度,长轴表示隐藏子结构所包含的节点数量。用户可以通过点击椭圆节点,展开隐藏的结构。主视图模块中的关键子结构如图 3 B3 所示。用户可以通过点击相关子结构,在节点连接视图中将高亮显示与关键子结构匹配的结构,若隐藏的结构中包含与关键子结构相匹配的结构,其对应的视觉提示也将高亮显示。

2.3　相似结构探索模块

相似结构探索模块主要由用户焦点结构(C2)、相似子结构(C3)、结构探索历史(C4)组成。

用户焦点结构为用户当前点击的关键子结构或框选的子结构。相似子结构部分呈现相似子结构匹配算法所提取的 top-k 个相似子结构。结构探索历史则将用户探索过的子结构相关信息进行保存,通过走马灯效果来回切换用户探索过的历史子结构。通过这样的方式保证用户能够及时切换前后关注的结构,从而对不同的子结构进行对比分析等。

同时,为便于用户更好地探索数据,设计了如图 3 C1 所示的控制面板。用户可以通过控制面板来控制主视图是否采用重要节点评估算法显示优化布局或者原始布局,这可帮助用户对比采用重要节点算法前后差异。对于布局方式可选择采用正交布局或者径向布局。除此之外,用户可使用框选工具在节点连接视图中框选感兴趣的子结构。

3　实　验

为验证该方法在对层次数据结构信息进行探索的有效性,采用用户评估以及案例研究来进行证明。

3.1　用户评估

实验目的:在文献[4]的研究中,证明了采用视觉提示的方式对于用户探索具有良好的帮助。但未说明是否只需部分的子结构采用视觉提示,就能加快数据探索。为证明所采用的节点重要性评估方法的有效性,将与文献[4]的 Tree Cues 方法进行用户对照实验。同时,为验证所提方法对于寻找相似子结构在精确度、完成时间上的有效性,将与文献[14]的 BarcodeTree 方法进行用户对照实验。

实验设计:用户对照实验方法参考文献[4,14]的方法进行展开。在实验之前,招募了 48 位志愿者,这些志愿者的年龄分布为 20 到 30 岁。为避免志愿者在完成不同实验过程中对数据产生了一定的熟悉度,将这 48 位志愿者分成两组来进行规避。第一组志愿者完成文中方法与 Tree Cues 的对照实验 A,第二组完成文中方法与文献[14]的用户对照实验 B。对照实验 A 设计主要针对节点级别上的探索,其实验内容参考文献[4]中的实验任务进行设计,以此保证在实验任务上的一致性。其任务详细设计为:TA1 标识层次结构中第二层(根节点为第一层)直接子节点数量 10 以内的节点。TA2 标识层次结构中第三层中直接子节点最多的节点。TA3 标识层次结构中第二层和第三层各 10 个直接子节点数量超过 10 个的节点。TA4 标识第三层的所有叶节点。

对照实验 B 的实验任务设计以子结构为主,其任务详细设计为:TB1 标识指定子结构在第二层和第五层的位置及数量。TB2 标识指定子结构,在第二层及第四层最相似的子结构。TB3 标识指定子结构,在第二层及第四层最相似的 10 个子结构

实验过程:对于对照实验 A 和 B,进行重复方差分析,将 a 设置为 0.05。实验 A 共进行 576 次,实验 B 共进行 432 次。在重复实验过程中,为避免用户对实验数据的熟悉程度产生依赖,每隔三到四天做一次实验,同时将对数据中每一层节点的位置在该层进行打乱布局顺序,进行重排。实验 A 中的每项任务都能获取准确值。因此,对于实验 A,只收集用户的完成时间。由于实验 B 中涉及寻找相似子结构,用户在寻找过程中是会存在错误情况的。因此在对照实验 B 中收集完成时间和错误率。

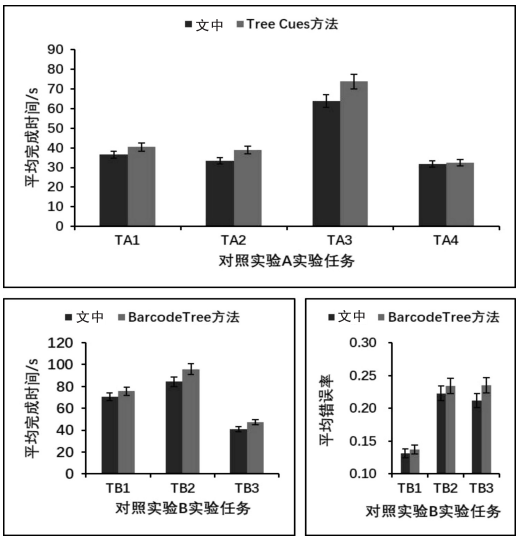

图 4　对照实验 A、B 实验结果

实验结果:图 4 显示了对照实验 A 完成实验的平均完成时间。表 1 中显示了各项任务完成时间的显著性检验结果。从表 1 的结果显示可知,在任务 TA1、TA2、TA3 的完成时间上,文中方法与文献[4]的方法具有显著性差异,说明在这三类任务的完成时间上文中方法是优于 Tree Cues 的。但对于类似任务 TA4 这种直接寻找叶节点,不需要其他额外的交互操作,两种方法在完成时间上并没有大的差异,也不具有显著性差异。这表明,文中采用的基于节点重要性对部分子结构采用视觉提示的方式,对于执行类似 TA1、TA2、TA3 这些任务时相比于 Tree Cues 中只采用视觉提示

的方法在完成时间上具有一定的优势。

对照实验 B 中采用文中方法与文献[14]中方法完成各项任务的时间和错误率,如图 4 所示。表 1 显示对照实验 B 中各项任务完成时间和错误率的显著性检验结果。从表 1 可知,在对照实验 B 中,完成时间上文中方法与文献[14]中的方法具有显著性差异。在任务错误率上,文中方法与文献[14]中的方法在 TB1 任务上不具有显著性差异,在 TB2 和 TB3 任务上具有显著性差异。该结果表明,文中方法在完成类似实验 B 中的任务时,在完成时间上相对于文献[14]中的方法更为高效。在完成任务错误率上的表现除 TB1 外具有显著差异。从对照实验 B 的结果可以得出,文中方法与文献[14]中的方法在精确度和完成时间总体上具有一定的优势。

表 1 对照实验

实验任务	显著性检验	错误率显著性检验
TA1	$F=10.401$ $p=0.001$	
TA2	$F=93.7$ $p<10^{-12}$	
TA3	$F=60.52$ $p<10^{-10}$	
TA4	$F=0.822$ $p=0.367$	
TB1	$F=79.16$ $p<10^{-12}$	$F=1.01$ $p=0.317$
TB2	$F=254.15$ $p<10^{-23}$	$F=4.27$ $p=0.042\,6$
TB3	$F=176.48$ $p<10^{-19}$	$F=17.56$ $p<10^{-4}$

A:各任务完成时间显著性检验与对照实验 B:各任务完成时间和错误率显著性检验

3.2 案例分析

该文设计了两个案例来说明该方法对于探索大型层次数据的有效性。

两个案例所使用的数据分别来源于中国政府统计网 2021 年成都市统计用区划和城乡划分代码和食安通网站上的农药残留限量数据。区划和城乡划分代码数据集是一个典型的层次数据集,详细反映了一个区域的行政层级划分情况。农药残留限量数据收集了 19 版和 14 版的数据。两个版本的层级划分都是以各类食物的类别进行详细划分,如水果类、蔬菜类等。

在两项案例分析中,分别招募了一位对案例分析数据陌生的志愿者参与案例研究中,避免志愿者的先验知识对实验产生影响。

案例一:以成都市的区划分代码数据为例进行案例分析。证明该系统能帮助用户通过交互式的方式从未知的大型层次数据中获取见解。

在对相似子结构进行探索的过程中,发现成都市的锦江区(图 5A)、成华区(图 5B)和青羊区(图 5C)的区划分的局部结构特征最为相似。同时这三个子结构的向量投影聚类结果也属于同一类簇。通过图 5 中显示可知,这三个区的子结构中大多由 4 到 8 个节点构成。通过这一现象说明这三个区的区划分具有一定的规律。查阅相关资料显示,这三个城区都为成都市的中心城区中的第一圈层。这从侧面反映了这三城区的区划分可能与其所处的地理位置、经济、人口等有一定联系。

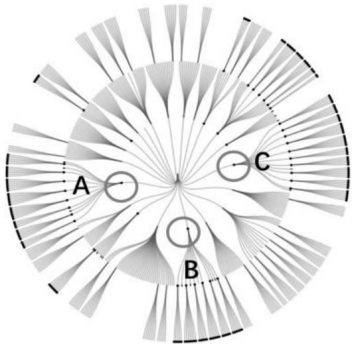

图 5 相似子结构探索对比分析

(A:锦江区 B:成华区 C:青羊区)

除此之外,用户通过探索关键子结构时发现了部分生僻的子结构,如图 6 中的 A、B 两类子结构在整个结构中远不如 C 这样的结构在整个区划分数据中分布的那么广泛。尤其 A 类结构的子节点数远多于 B、C 两类结构。为探知出现这一现象的原因,与用户一起查阅了相关资料。资料显示出现 A 类结构的新都区其面积在成都市排名第三,这一数据间接反映出 A 类结构的出现可能因其地理因素影响较大。出现 B 类结构的城区为金牛区和武侯区作为一圈城的城区,出现该情况可能跟其他因素有关如人口等。

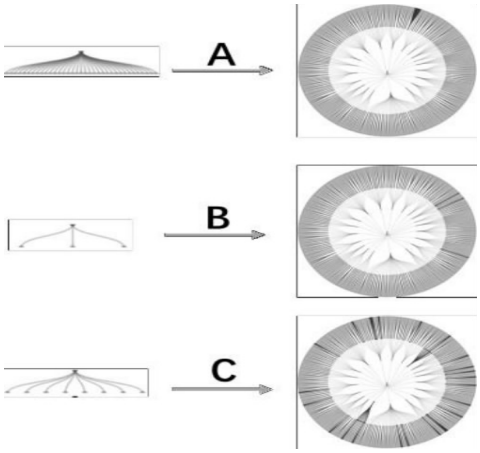

图 6 A、B、C 三类子结构在整个层次结构中的分布情况

案例二:探索不同版本的层次数据之间的差异,可帮助用户快速了解版本之间的演变过程,同时也能协助对错误进行追踪[15]等。选择 19 版和 14 版农药残留限量数据证明该方法可用于辅助探索具有版本迭代地层次数据的结构差异性。

为探索两个版本之间的差异,志愿者首先对比两版本的数据的概览模块。对比发现,在 19 版中第四层总的节点数由 14 版的 170 多增加到 200。其增加节点类型以叶节点为主。这说明,在 19 版本中引入了更多的测定对象。然后又进行了关键子结构对比,发现在两个版本中,其关键子结构的数量都为 7 种,但关键子结构只有 4 个保持完全相同(见图 7)。志愿者接着对比了关键子结构在原始布局上的整体分布。在对比的过程中发现,19 版的第四层节点中删除了部分 14 版中的叶节点,如在谷物这一大类下 19 版删除了 14 版中的一个叶节点(其他麦类)。这一细微差异相对于结构差异变化明显的来说是很难发现的(如油料油脂分类的子类在 19 版由四大分类变为两大分类)。除对比两个相同子结构在整体上的分布外,志愿者在对比两个相似的子结构时,同样发现了类似上述的差异情况。

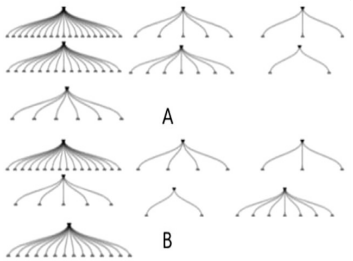

图 7　关键子结构对比农药残留限量数据 14 版(A)与 19 版(B)

4　结束语

与现有的方法相比,提出采用重要节点评估算法来评估节点的重要性,根据重要性来决定需视觉编码的节点,这能保留更多的拓扑结构信息避免过多的结构信息被视觉编码隐藏,从而提高了探索效率。同时,采用提取关键子结构对层次结构的拓扑特征进行概括和图嵌入算法对节点进行向量化方法相结合,提升了用户探索拓扑结构信息的效率与准确率。实验案例研究表明,该方法不仅支持单一层次结构数据的拓扑结构信息探索,还适用于具有版本迭代的层次结构信息的对比分析。

由于层次数据不仅蕴含着结构信息,同时各节点中具有属性信息,在未来的工作中将研究或提出一种合理的方法将属性信息加入当前探索框架中,来帮助用户更全面地探索层次数据中的结构信息和属性信息。

参考文献:

[1] CHEN C. Information visualization and virtual environments [M]. London:Springer-Verlag,1999:223.

[2] NGUYEN Q V,HUANG M L. Enc Con:an approach to constructing interactive visualization of large hierarchical data [J]. Information Visualization,2005,4(1):1-21.

[3] REINGOLD E M,TILFORD J S. Tidier drawing of trees [J]. IEEE Transactions on Software Engineering,1981,7(2):223-228.

[4] NGUYEN Q V,ARNESS D,SANDERSON C J,et al. Enabling effective tree exploration using visual cues[J]. Journal of Visual Languages & Computing,2018,47:44-61.

[5] HEER J,CARD S K. DOITrees revisited:scalable,space-constrained visualization of hierarchical data[C]//Proceedings of the working conference on advanced visual interfaces. Gallipoli:DBLP,2004.

[6] 张　玫. 节点重要性评估及其在城市公交网络中的应用[D]. 石家庄:河北师范大学,2016.

[7] SHNEIDERMAN B. The eyes have it:a task by data type taxonomy for information visualizations[C]//Proceedings of the IEEE symposium on visual languages. Boulder:IEEE,1996:336-343.

[8] CARLETTI V, FOGGIA P, VENTO M. VF2 plus:an improved version of VF2 for biological graphs[C]//Proceedings of the int'l workshop on graph-based representations in pattern recognition. Beijing: Springer - Verlag, 2015: 168 - 177.

[9] CARLETTI V,FOGGIA P,SAGGESE A,et al. Introducing VF3:a new algorithm for subgraph isomorphism[C]//Proceedings of the int'l workshop on graph-based representations in pattern recognition. Anacapri: Springer - Verlag, 2017:128-130.

[10] 潘嘉铖,韩东明,郭方舟,等. 面向比特币交易网络的拓扑结构可视探索方法[J]. 软件学报,2019,30(10):3017-3025.

[11] 李　珍. 基于图表示学习的子图同构约束求解技术研究[D]. 桂林:桂林电子科技大学,2021.

[12] DONNAT C,ZITNIK M,HALLAC D,et al. Learning structural node embeddings via diffusion wavelets[J]. arXiv:1710.10321,2018.

[13] SHERVASHIDZE N,SCHWEITZER P,LEEUWEN E J,et al. Weisfeiler-Lehman graph kernels[J]. Journal of Machine Learning Research,2011,12(9):2539-2561.

[14] LI Guozheng,ZHANG Yu,DONG Yu,et al. BarcodeTree:scalable comparison of multiple hierarchies[J]. IEEE Transactions on Visualization and Computer Graphics,2020,26(1):1022-1032.

[15] LESCHKE T R,NICHOLAS C. Change-link 2.0:a digital forensic tool for visualizing changes to shadow volume data [C]//Proceedings of the 10th workshop on visualization for cyber security. Atlanta:ACM Press,2013:17-24.