文章编号:1001-2486(2011)01-0115-05

Parallel Sets的改进及其在全球恐怖袭击数据分析中的应用*

肖卫东,周 城,孙 扬,葛 斌,汤大权
(国防科技大学 C^4ISR 技术国防科技重点实验室,湖南 长沙 410073)

摘 要:随着恐怖主义愈演愈烈,"反恐"成为当今世界各国军事安全部门的中心任务。使用分类型可视化工具 Parallel Sets 分析国际恐怖主义数据库中多属性分类值间的关系,揭示数据库中的隐性信息,并针对 Parallel Sets 任意排列分类值产生较多交叉的不足,提出带降势的启发式分类值布局算法,自动优化分类值布局顺序,减轻视图中的可视混乱,降势策略可以减少参与计算的分类值数目。实验结果表明,改进的 Parallel Sets 可清晰展现国际恐怖主义数据库中各分类值间的关联,从而辅助用户分析不同恐怖组织的行为特征等信息;带降势的启发式分类值布局算法简单高效,适用于数据量较大、分类值较多的数据集。

关键词:平行集;边交叉问题;恐怖主义;可视化分析;降势
中图分类号:TP391 **文献标识码**:A

Improvement of Parallel Sets and Its Application in Analyzing Global Terrorism Database

XIAO Wei-dong, ZHOU Cheng, SUN Yang, GE Bin, TANG Da-quan
(C^4ISR Technology Key LAB, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: With terrorism aggravating, anti-terrorism has been a main task for national military security departments around the world. The current study utilized categorical data visualization, Parallel Sets, to analyze the relations among the multi-categories in Global Terrorism Database, aimed to uncover the implicit information within the data set. To alleviate the deficiency of excessive edge crossing brought by random layout of categorical values, the research proposed a heuristic layout algorithm based on average heuristic with cardinality reduction, which optimized the layout order of categories and the visual clutter is eased so that the cardinality reduction strategies can reduce the numbers of categories involved in computation. The experimental results demonstrate that the improved parallel sets can clearly express the association among the multi-categories in Global Terrorism Database, thereby assist users in analyzing the information of various terrorist organizations, such as the behavior characteristics. Furthermore, the average-based heuristic with cardinality reduction is simple and highly efficient, which is suitable for large data sets with many categorical attributes.

Key words: parallel sets; edge crossing; terrorism; visual analytics; cardinality reduction

自"911事件"之后,恐怖主义成为世界最为关注的社会问题之一,"反恐"也就成为不少国家军事及安全部门的中心任务。在过去几年中,为能应用信息技术有效支持"反恐"工作,维护世界和平,越来越多的专家学者加入到研究全世界恐怖分子活动行为的行列中。但是,由于大多数有关恐怖主义的信息都以不同形式分散于不同数据集中,未形成一个统一完整的共享数据环境,难以有效支持相关人员全面系统地进行研究分析,发掘其中的真相、特征、规律、趋势及关联信息。而由美国马里兰大学恐怖主义研究应对协会管理的免费开放的全球恐怖主义数据库(Global Terrorism Database,GTD),详细记录了1970至今世界范围内8万多起恐怖袭击事件,很好地解决了这一共享数据环境的问题[1]。

随着GTD数据规模的增加,研究人员愈发感到缺乏有效的数据分析理解工具辅助其发掘隐含信息、验证相关假设及解释对应结论,因此,各国学者开始尝试使用信息可视化工具分析GTD,这样不仅能够有效利用计算机图形表现抽象数据,而且可以借助视觉增强用户对非物理抽象信息的认知。如Joonghoon提出的GTD Explorer根据不同的分类标准统计每年恐怖袭击发生数量,而后使用堆积面积图(stack charts)[2-3]进行可视化,并设

* 收稿日期:2010-05-18
基金项目:国家自然科学基金资助项目(60903225);国防科技大学优秀研究生创新基金资助项目(B080503)
作者简介:(1968—),男,教授,博士。

计实现了一套特色而友好的交互方法,用于分析一类或一组袭击事件随时间的变化趋势[4];杨育彬等结合社会网络可视化分析和数据挖掘的理论与方法,并利用社会网络分析法对恐怖袭击事件各要素节点间关系进行量化表征及分析,从而实现了对恐怖组织间的活动模式及发展特点等内在规律的挖掘及解释[5];Wang 等将多种可视化方法及视图有效地进行整合集成,以 5W(who, what, where, when, why)理论为基础设计了交互式分析系统,通过可视提供恐怖袭击的发生时间、发生地点、相关恐怖组织间的相互关联信息及袭击方法和手段,辅助相关人员迅速查找分析事件原因[6],等等。

虽然上述工具可以从时序演化、社会网络、地理信息等方面对 GTD 进行可视分析,但无法有效分析在 GTD 中占大多数的分类型属性,如目标类型、袭击手段、武器类型、袭击区域等。本文试图使用分类型数据可视化方法对 GTD 中多属性分类值间的关系进行可视分析,以发掘不同恐怖组织的行为特征及针对特定目标的袭击特点等隐性信息。Bendix 等提出的 Parallel Sets[7] 对较大规模分类型数据集的适用性相对较好。但是,Parallel Sets 在每条平行轴上对分类值进行任意排列,会使对应二平行轴间的关联平行四边形产生较多交叠,从而引起视图的可视化混乱,加重分析人员的认知负担,阻碍其清晰高效地解读各分类值间的关系,虽然 Parallel Sets 实现了对分类值线段的拖拉交互以改变其排列顺序,但是若只通过手动操作将视图调整清晰是一项十分繁杂的任务。因此,针对上述 Parallel Sets 的缺陷,本文提出带降势的启发式分类值布局 CLEARCR 算法(Categories Layout basEd on Average heuRistic with Cardinality Reduction)自动优化各平行轴上分类值的布局,以减少相关平行四边形产生的交叠数量,并应用改进的 Parallel Sets 对 GTD 进行可视分析,实验结果表明,改进的 Parallel Sets 可以清晰展现 GTD 中各分类值间的关联信息,从而有效地辅助相关人员分析 GTD 中隐含的恐怖组织活动特点等信息;CLEARCR 算法可明显减少 Parallel Sets 中关联平行四边形产生的交叠数量,且复杂度较低,适用于数据量较大、分类值较多的数据集。

1 CLEARCR 算法

为明确阐述问题,我们对 Parallel Sets 可视化视图进行抽象,将每条平行轴上代表分类值的线段用节点代替,将连接不同分类值的平行四边形用直线代替,层间连线说明数据集中存在包含对应两分类值的记录,且连线带有权重信息。这样 Parallel Sets 视图就转换为一层次网络图,如图 1 所示,该类层次网络图可形式化描述为 $G(L_1, L_2, L_3, \cdots L_n, E)$,其中,$L_i$ 表示同一层次的一组有序节点,$E \subseteq (\bigcup(L_i \times L_{i+1}))$代表节点组 L_{i+1} 与其上层节点 L_i 连线的集合,集合中的任一元素 uw 表示节点 $u(u \in L_i)$ 与节点 $v(v \in L_{i+1})$ 的连线。

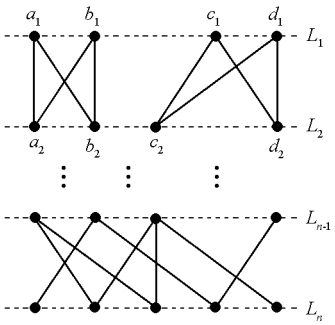

图 1 层次网络图
Fig. 1 Hierarchical network

经研究发现,两条平行轴之间的边交叉数量与对应层节点排列顺序密切相关,由于图 1 中平行轴上的节点是无序的,因此可以通过计算最佳节点排列顺序,减少因边交叉引起的可视混乱,优化可视化视图,我们称之为边交叉问题,即对于一层次网络图 $G(L_1, L_2, L_3, \cdots L_n, E)$,求解 L_i 层对应节点的顺序 $order_i$ 使得 $cross(G, order_1, order_2, \cdots, order_n)$最小。

Eades 曾提出针对二部图(Bipartite Graphs, $G(L_1, L_2, E)$)的中位数启发式算法[8],其思路是:首先给定第一层节点的位置(顺序),然后计算与第二层每一节点相连接的第一层节点的中位点,并将该中位点的位置作为第二层对应节点的位置,最后根据第二层全部节点的位置对其进行排序绘制。Eades 的算法只解决了二部图的绘制,无法直接处理如图 1 所示的多层次网络图,因此,我们提出适合多层次网络图的绘制算法,其思路是将一多层次网络图从上到下每两层依次迭代使用 Eades 算法,当到达底部时反向迭代,直至条件满足为止。而且我们认为采用平均值能减少取得相同坐标值的节点数目,减少了随机排序的可能性,因而能取得比中位数启发式算法更好的效果,故本文采用平均值启发式算法。但是,随着层次网络图中各层次节点数量的增加,导致该算法的内存和时间耗费问题越来越严重。因此,本文提出

利用节点聚类算法来减少参与排序的节点数量(降势),原来的若干个节点由聚类产生的新节点所代替。考虑到边交叉是由节点的相邻关系产生,本文在定义邻接向量的基础上提出一基于向量比较的算法来实现聚类过程。为能形式化描述 CLEARCR 算法及聚类算法,在图 1 的基础上给出如下定义:

定义 1　对任意 $v(v\in L_{i+1})$使用 N_v 表示集合$\{u:uv\in E,u\in L_i\}$,即 v 的"上"层相邻节点构成的集合,称集合基数$|N_v|$为 v 的相邻度。

定义 2　对任意 $v(v\in L_{i+1})$定义邻接向量 $\boldsymbol{X}_v=[x_{vu_1},x_{vu_2},\cdots,x_{vu_n}]$,当 $u_i\in N_v$ 时,$x_{vu_i}=1$,否则 $x_{vu_i}=0$。例如,在图 1 中,$\boldsymbol{X}_{a2}=[1,1,0,0]$,$\boldsymbol{X}_{b2}=[1,1,0,0]$,$\boldsymbol{X}_{c2}=[0,0,1,1]$,$\boldsymbol{X}_{d2}=[0,0,1,1]$。

定义 3　对任意两个节点的向量 $\boldsymbol{X}_v=[x_{vu_1},x_{vu_2},\cdots,x_{vu_n}]$,$\boldsymbol{X}_w=[x_{wu_1},x_{wu_2},\cdots,x_{wu_n}]$,定义初始距离 $d_{v,w}=0$,当 i 从 1 到 n 时,依次比较两向量的对应分量,当 $x_{vui}\neq x_{wui}$时,距离 $d_{v,w}$ 加 1。另外,用户可设置一个阈值 h,当 $d_{v,w}\leqslant h$ 时,对应两节点聚为一类。例如,在图 1 中,$d_{a2,b2}=0$,$d_{a2,c2}=4$,$d_{a2,d2}=4$,$d_{c2,d2}=0$,假如现定义阈值 $h=0$,则节点 a_2,b_2 聚为一类,c_2,d_2 聚为一类。

定义 4　根据定义 3 中的聚类算法,可以将每层节点进行聚类,每一类 C_j 抽象为新的节点 p_j。如果其它节点 w 与 C_j 中任一节点有相邻关系,那么在抽象出的层次图中,该节点 w 与 p_j 也存在相邻关系。

定义 5　CLEARCR 算法中任意节点的坐标值由其"上"层相邻节点的坐标值决定,假定节点 $u\in L_i$ 的坐标值为 $X_i(u)$,节点 $v\in L_{i+1}$的相邻节点为 $\boldsymbol{N}_v=\{u_1,u_2,u_3,\cdots,u_j\}$,那么,$v$ 的坐标值 $X_{i+1}(v)=\sum_{i=1}^{j}X_i(u)/j=\text{avg}(N_v)$。若 v 无"上"层相邻节点,则定义 $X_{i+1}(v)=0$。

为能更清晰地说明算法过程,图 2 给出了 CLEARCR 算法,并以图 3 所示的层次网络图为例解释算法执行过程,这里阈值 h 取默认值 0。首先任意给定一层次网络图(图 3(a));其次通过对第二层聚类形成新的层次网络图(图 3(b));第三步通过对第三层聚类形成另一新的层次网络图,任意给定第一层节点的排列顺序并分配坐标值,并根据相邻节点的平均数计算第二层节点的坐标值(图 3(c));第四步根据坐标值排序第二层节点并重新分配坐标值,而后计算第三层节点的坐标

CLEARCR 算法

(1)对给定的一层次网络图 G,通过逐层聚类抽象,最后抽象形成新的层次网络图 G';
(2)给定图 G' 初始 L_1 层节点 $u_1,u_2,\cdots,u_k$ 的顺序 $order_1$,并为 u_l 分配坐标值 $X_1(u_l)=l$;
(3)根据 L_i 层节点的坐标值计算 L_{i+1}层全部节点 v 的坐标值 $X_{i+1}(v)=\text{avg}(N_v)$;
(4)以坐标值 $X_{i+1}(v)$为依据对 v 进行层内排序,得到 $order_{i+1}$,其中,若两节点坐标值相等,则将相邻度小的节点排在前面,相邻度相等时,这些节点按原来前后顺序排列;
(5)根据 $order_{i+1}$重新为 v_l 分配坐标值 $X_{i+1}(v_l)=l$;
(6)重复执行(3)~(5),当 $i=n$ 时,改变迭代方向,即:L_n 变为 L_1,L_{n-1}变为 L_2,L_1 变为 L_n;
(7)重复步骤(3)~(6),直至可视化视图中不存在交叉,或者交叉总数不再变化,或者达到初始设定的终止迭代条件(如交叉数量或执行步骤等);
(8)将由聚类形成的节点逐层替换为原来节点,并恢复原来的相邻关系。

图 2　CLEARCR 算法

Fig.2　CLEARCR algorithm

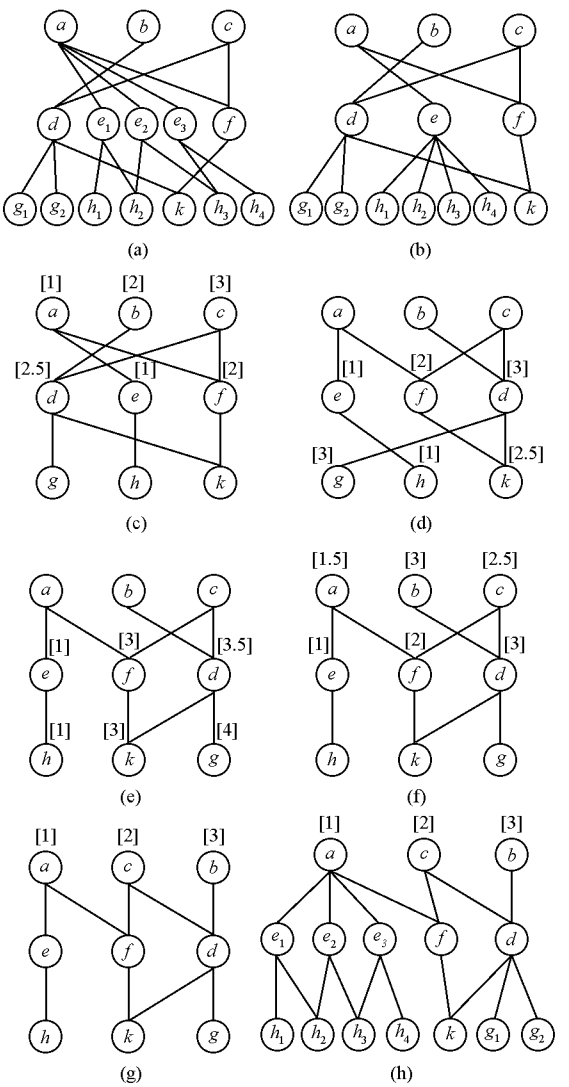

图 3　CLEARCR 算法执行顺序图

Fig.3　Process of CLEARCR algorithm

值(图 3(d));第五步根据坐标值排序第三层节点并重新分配坐标值,同时转换迭代方向,计算第二层节点的坐标值(图 3(e));第六步根据坐标值排序第二层节点并重新分配坐标值,而后计算第一层节点的坐标值(图 3(f));第七步根据坐标值排序第一层节点,这时候层次网络图中已不存在边交叉,每一层得到合理的分类值排列顺序(图 3(g));第八步将由聚类形成的节点替换为原来的节点,并恢复原先相邻关系,算法终止。

由于 CLEARCR 算法属于启发式算法,无法保证其总能输出边交叉数量最少的分类值布局结果,但是,实验结果表明该算法在多数情况下都能给出边交叉较少的分类值排列顺序。

3 实验结果

通过使用改进的 Parallel Sets 可视化分析全球恐怖袭击数据,发掘其中的隐含信息,验证 Parallel Sets 对 GTD 中分类属性分析的有效性,并通过对比改进前后 Parallel Sets 可视化结果验证 CLEARCR 算法的重要性及正确性,最后通过分析实验数据证明 CLEARCR 算法减少边交叉的能力及其较好的时间性能。

原型系统使用 Eclipse 基于 SWT 控件编写实现,操作系统平台 Microsoft Windows XP,机器配置为:Intel T5870,2GB 内存,300GB 硬盘。实验中使用的全球恐怖袭击数据集包含 2006 年 947 条恐怖事件的记录,选择了 5 个分类型属性进行分析,分别是 PERPETRATOR(Taliban, Abu Sayyaf Group, Hamas, Liberation Tigers of Tamil Eelam, Salafist Group for Preaching and Fighting 及 Naxalites), TARGET TYPE (Police, Business, Military, Transportation), REGION(South Asia, Middle East & North Africa, Southeast Asia), ATTACK TYPE (Bombing/Explosion, Armed Assault, Facility/Infrastructure Attack, Hostage Taking(Kidnapping)), WEAPON TYPE (Explosives/Bombs/Dynamite, Firearms, Melee, Armed Assault),括号内表示选取的分类值。数据来源于 http://www.start.umd.edu/gtd/。

3.1 使用改进 Parallel Sets 可视化分析 GTD

采用改进的 Parallel Sets 可视化上述恐怖数据集,结果如图 4 所示。由图 4 可以清晰直观地获取各恐怖组织的活动特点,如 Taliban 的袭击目标遍及警察、商业及军队,大多装备炸药及轻武器,主要以爆炸及武装袭击为手段,但其作案地域限于南亚,而同在南亚作案的 Naxalites 只使用炸弹袭击警察,另外同在南亚作案的 Tamil Eelam 猛虎组织主要利用炸弹袭击军事设施,而 Salafist Group 集中在中东对警察和交通设施进行炸弹袭击;通过分析恐怖组织的活动范围,发现恐怖组织的地域性极强,不会轻易转移根据地;通过分析袭击目标类型及武器类型,发现袭击警察大多以轻武器为主,而袭击军队则以炸药为主;通过观察发现视图中没有 Kidnapping 和 Melee 项,说明所选取的恐怖组织未涉足绑架事件,而且他们不会赤手空拳地发动袭击。获取的信息可以有力地支持军事及安全部门采取针对性措施积极防御恐怖袭击,并且可以根据袭击特点推断执行组织。通过交互操作可进一步分析,发掘更多 GTD 中的隐性信息,在此限于篇幅不再一一赘述。

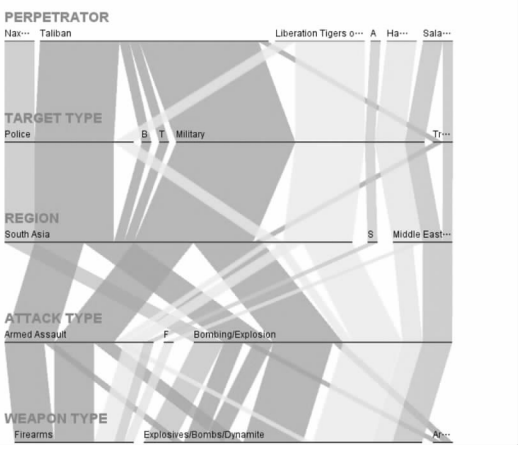

图 4 改进的 Parallel Sets 可视化效果图
Fig.4 Visualization by improved Parallel Sets

图 5 是未使用 CLEARCR 算法任意排列分类值的可视化效果图,虽然也可模糊得到上述部分结论,但是整个视图由于存在很多交叉显得非常混乱,这在无形中给用户造成了相当程度认知障碍,当然,我们可以通过交互操作调整出较清晰的可视化视图,但是该过程相当的繁琐复杂,需要耗

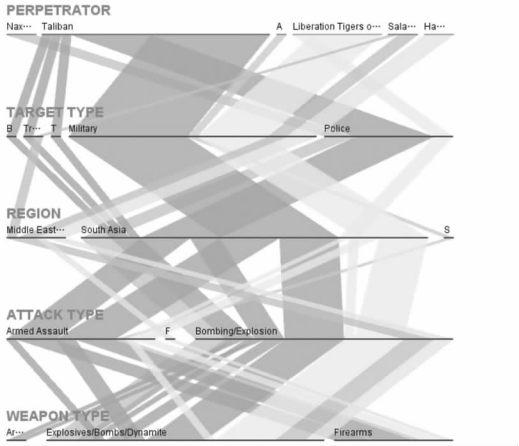

图 5 任意排列的 Parallel Sets 可视化效果图
Fig.5 Visualization by arbitrary Parallel Sets

费大量时间,并且不确定性较大。

3.2　CLEARCR 算法效果及性能分析

为验证 CLEARCR 算法减少边交叉的效用及时间复杂度,通过调节 2006 年 GTD 数据集的记录条数、属性数和分类值总数分别进行可视化,试图发现交叉数与这三个特征的关系,其中分类值尽可能平均分布在 Parallel Sets 的每一属性轴上。

CLEARCR 算法对不同数据集的作用效果如表 1 所示,运行效率如图 6 所示。从表 1 中可知,随着数据集记录条数、属性数和分类值总数的增加,任意排列和 CLEARCR 算法排列产生的交叉数量都会呈现不同程度的增加,尤其受分类值总数的影响最大,而与记录条数及属性数的关系不甚密切;相比于任意排列,CLEARCR 算法可明显减少交叉数量,大概可减少原交叉数量的 2/3 左右。由图 6 可知,CLEARCR 算法的运行效率与数据集的记录条数和属性数的关系不大,主要由分类值总数决定,总体看来算法的时间复杂度较低,在用户可以承受的范围之内。

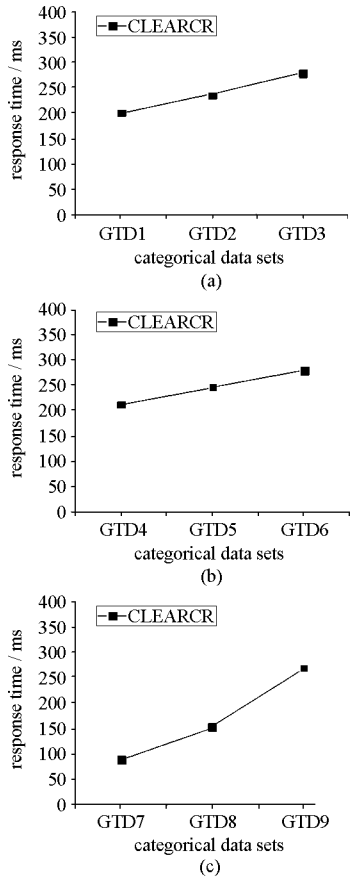

图 6　CLEARCR 算法性能分析图

Fig.6　Performance of CLEARCR algorithm

表 1　CLEARCR 算法效果对比分析表

Tab.1　Effect comparison of CLEARCR algorithm

数据集	记录条数	属性数	分类值总数	任意排列产生交叉数	CLEARCR 算法排列产生交叉数
GTD1	474	5	21	148	49
GTD2	711	5	21	154	54
GTD3	947	5	21	161	58
GTD4	947	3	21	143	47
GTD5	947	4	21	151	52
GTD6	947	5	21	161	58
GTD7	947	5	10	11	3
GTD8	947	5	15	38	12
GTD9	947	5	20	151	51

4　结论与进一步工作

本文针对因 Parallel Sets 任意排列分类值产生的边交叉问题,提出 CLEARCR 算法自动优化分类值布局,减轻由边交叉引起的可视混乱现象。并采用改进的 Parallel Sets 可视化分析 GTD 中多个属性及多属性分类值间的关系,辅助用户发掘恐怖数据集中的隐性信息。实验结果表明,改进 Parallel Sets 适合处理数据量较大的恐怖数据集,可视化视图易于理解,能直观准确地揭示 GTD 中隐含的恐怖组织活动规律等信息;CLEARCR 算法简单,能有效减少可视化视图中的边交叉数量,时间复杂度较低,适用于数据量较大、分类值较多的数据集合。

在后续工作中,我们将进一步改进 CLEARCR 算法,优化其减轻可视混乱的效果,设计更加友好的交互操作辅助用户进行可视化分析,进行更加正式严格的用户评测实验证明改进的 Parallel Sets 在对 GTD 分析任务中的优越性。

参 考 文 献:

[1] LaFree G, Dugan L, Fogg H V, et al. Building a Global Terrorism Database[R]. Report to the US Department of Justice, 2006.

[2] Wattenberg M. Baby Names, Visualization, and Social Data Analysis [C]//Proceedings of IEEE Symposium on Information Visualization, 2005.

[3] Havre S, Hetzler E, Whitney P, et al. ThemeRiver: Visualizing Thematic Changes in Large Document Collections [J]. IEEE Transactions on Visualization and Computer Graphics, 2002, 8 (1): 9 – 20.

[4] Lee J. Exploring Global Terrorism Data: A Web-based Visualization of Temporal Data[J]. Crossroads, 2008, 15(2): 7 – 14.

[5] 杨育彬,李宁, 张瑶. 基于社会网络可视化分析的数据挖掘[J]. 软件学报, 2008, 19(8): 1980 – 1994.

[6] Wang X, Miller E, Smarick K, et al. Investigative Visual Analysis of Global Terrorism[J]. Computer Graphics Forum, 2008, 27(3): 919 – 926.

[7] Bendix F, Kosara R, Hauser H. Parallel Sets: Visual Analysis of Categorical Data [C]//Proceedings of IEEE Symposium on Information Visualization, 2005.

[8] Eades P. Edge Crossings in Drawings of Bipartite Graphs[J]. Algorithmica, 1994, 11: 379 – 403.