多维网络论坛数据的层次可视化

许彦如[1] 王长波[1,2] 刘玉华[1] 章群燕[1]
(华东师范大学软件学院 上海 200062)[1] (浙江大学 CAD&CG 国家重点实验室 杭州 310058)[2]

摘 要 多维数据的可视化可以帮助人们在有限的时间内快速理解和分析海量数据集。网络论坛数据具有复杂性和大规模性,无法采用简单的图形显示出隐含的规律。针对多维海量论坛数据,首次分析了论坛数据的特点,然后提出了一种层次的数据组织方式,进而针对论坛中的主题时变分布、热门主题转换、作者回复关系等多种数据信息,采用平行坐标、曲面关系图、层次映射图等多种形式进行海量论坛数据的互动可视化。最后以实际论坛数据为例,给出可视化结果并进行分析。

关键词 论坛数据,多维信息,层次可视化
中图法分类号 TP391 **文献标识码** A

Hierarchical Visualization of Multi-dimensional Forum Data

XU Yan-ru[1] WANG Chang-bo[1,2] LIU Yu-hua[1] ZHANG Qun-yan[1]
(Software Engineering Institute, East China Normal University, Shanghai 200062, China)[1]
(State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou 310058, China)[2]

Abstract Visualization of multi-dimensional data is an effective method for helping people understand and analyse multivariate data sets. Forums in Internet have taken a more and more important role in everyone's life. Visualizing forum data will help analyze the pattern of people's participation in forums. This paper first analyzed the feature of forum data such as closely relation with time, mass data size, multi-level data structure, multiple attributes and complex relation within the data. We proposed three methods, such as parallel coordinate, hierarchical map and curve map to visualize the different datasets in network forum. At last, we took data in a real forum for example and showed the results of visualization.

Keywords Forum data, Mass data, Visualization of hierarchies

1 引言

随着网络技术的发展和网络规模的不断扩大,网络已经成为人们日常生活中不可缺少的一部分,越来越多的人参与到网络这个虚拟世界中,各种论坛、兴趣小组随之兴起,丰富人们的生活。由于论坛数据涉及到海量的主题和内容信息,对论坛数据进行可视化分析,可以大大方便论坛的管理和组织,使论坛管理者和其他研究人员根据可视化结果做出决策,也可以使人们快速了解论坛信息。

目前,国内外已经有一些关于论坛这一类网络数据的可视化工作。Smith 和 Fiore[1] 使用多种组件可视化新闻组中的讨论区层次结构以及讨论过程中用户参与的模式,从不同侧面反映这些信息。Paolillo[5] 从某一个讨论区中抽取出词频列表,从讨论区使用语言、讨论的主题以及用户参与的频度等方面可视化用户网络结构来帮助用户定位到感兴趣的话题。Viégas 和 Fiore[2] 采用两种工具 Newsgroup Crowds 和 Author Lines 分别可视化在某个讨论区中作者的受欢迎程度以及某个人在讨论区中发言的情况。Welser 等[3] 基于用户网络新闻组,分析讨论区中不同人扮演角色的特征,并重点针对"answer people"这类角色的 3 种特征,从不同角度可视化分析属于"answer people"的人。Fisher 和 Marc Smith[4] 利用用户网络新闻组中的作者回复可视化作者间的关系网络,区分不同作者在新闻组中扮演的不同角色。Heer 和 Boyd[6] 开发了系统 Vizster,用图的方式可视化在线社会网络。

除了基于新闻组的可视化还有一些工作是关于电子邮件数据的可视化。Fernanda 等[7] 开发了系统 Themail,利用电子邮件存档信息可视化作者间的关系,并从电子邮件的内容中找出一个能描述作者和其他人之间讨论内容的单词以及这个单词随着时间关系如何改变。Kerr[8] 介绍了 Thread Arcs,它是一种可视化工具,将电子邮件的时间信息和电子邮件的回复的树形结构结合起来,方便人们快速查看邮件内容。Rohall 等[10] 介绍了一种可视化大规模电子邮件存档工具的设计和开发细节,以帮助人们发现隐藏在电子邮件中的信息。Venolia 等[11] 给出了一种可视化 Email 存档的方法,将传统

到稿日期:2010-03-10 返修日期:2010-06-01 本文受国家 973 项目(2010CB731406),国家自然科学基金项目(60603076),上海市青年科技启明星计划项目(08QA14025),上海市科委自然科学基金项目(07ZR14035),浙江大学 CAD&CG 国家重点实验室开放基金项目(A1008)资助。
许彦如(1987—),女,硕士生,主要研究方向为可视化分析;王长波(1976—),男,博士,副教授,主要研究方向为真实感图形、可视分析、虚拟现实等,E-mail:cbwangcg@gmail.com(通信作者)。

非人性化的 Email 界面转化为用户友好的图形化界面。

国内学者目前对网络数据的可视化方面的研究还较少，裴新等[12]采用极大 k-plex 发现算法来挖掘网络社区中的简历信息；黄雄伟等[13]结合基本的 Web 数据挖掘算法，提出了 Web 数据挖掘可视化实现框架；蔡磊等[15]提出了一种适用于空间数据的四叉树金字塔的分层分块方案，以实现海量空间数据的可视化；杨育彬等[16]利用社会网络分析的方法对社会现象进行分析和可视化。

以上这些可视化方法，要么从单个人的角度可视化网络论坛数据，以方便人们查找邮件或新闻的历史信息，没有从整个网络数据集的角度来可视化分析其中存在的规律；要么虽然从整个网络的角度来分析，但是其数据量只有几百个，不能涵盖大规模或者海量的论坛数据。本文基于从网络论坛上抽取出来的数据，对海量网络数据集的回帖时变分布、作者关系、热点转移等信息进行可视化分析，探讨了大规模网络论坛数据的可视化方法。

2 论坛数据可视化

2.1 论坛数据的特点

网络论坛数据与其他领域可视化需求的数据集相比，有下面几个特点：(1)数据与时间密切相关。论坛实时发生着变化，用户在某个时刻注册成为论坛用户，又在某个时刻发表了帖子，在某个时刻回复帖子，这些数据因为有时间属性才有意义。(2)数据量大。一般论坛都分为好多板块，每个板块中的帖子量高达几千至上万个，热门帖子可能有上千人次的回复，每次回复中又有回复的时间和发表这次回复的作者的信息，论坛中有上万乃至百万人的参与，所以总的数据量很大。(3)数据层次化。在论坛中，信息以层次的方式组织起来，一般论坛的第一层是各个版块，第二层便是每个版块下面所属的帖子，帖子又是由以跟帖者的回复形成的最小单位构成的。(4)数据维数多，复杂程度高。论坛中的数据，从不同角度可以观察出不同的维度信息。从作者的角度，作者在论坛中的注册时间、发帖回帖时间、作者积分、作者所在地区等等都构成了作者的属性。从帖子的角度，帖子的发帖时间、发帖人、帖子的长度、持续的时间等等都构成了帖子的属性，与此相类似的还有论坛的属性。而论坛中作者之间回复的关系，构成了一个非常复杂的网络。

2.2 论坛数据的可视化方式

根据如上所述论坛数据的几个特点可以看出，对论坛数据可视化不能采用对一般数据可视化的方法，一方面由于论坛数据量很大，使得我们无法一次性全部显示所有信息，一次性显示将可能导致数据显示的混乱与重叠，不能达到期望的可视化效果。而且，一次性显示需要计算的数据量比较大，可能导致系统执行时间相对较慢。

在这里，我们充分利用论坛数据的层次性以及与时间相关的特性，将论坛的数据按层次组织起来进行层次可视化。初始时，仅仅对第一层数据的统计信息进行可视化显示，只有在用户请求时，才对第二层的统计信息进行可视化显示。同样，在可视化出第二层的统计信息时，只有在用户请求时才可视化显示第三层统计信息。类似地，在可视化显示最后一层统计数据时，只有在用户请求下，才可视化显示出具体的数据。

这里我们关心的数据主要包括 3 个层面：(1)论坛中不同时间、不同地域、不同板块的帖子分布及情况分析，以便对这个论坛中的数据进行管理；(2)论坛中的热点话题数据，在不同阶段网络论坛的热门话题是随着时间变化的；(3)论坛中参与者的回帖讨论关系，选取论坛中参与者的回复关系，可视化出论坛中不同人的参与情况。这些信息既涉及到大范围的全局信息，也涉及到参与人的地域、回帖等信息，其可视化有助于多维论坛数据的可视化管理和分析。

3 海量信息的层次化组织

3.1 数据的层次化组织方式

要进行海量论坛信息的层次可视化，首先要以层次化的方式组织数据。对于要可视化的论坛数据，我们以这样的方式组织：第一层将整个数据集按照版块的不同划分为几个数据集，接着将划分后的数据集按照发帖时间的不同划分为低一级的数据集，最后将划分后的数据集按照发帖人所属区域的不同划分为更低一级的数据集，如图 1 所示。

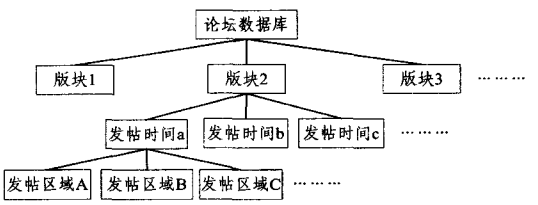


图 1 论坛数据集的划分

3.2 可视化映射

分层组织好数据以后，就要将数据映射到具体的表现形式。数据被划分为 4 层，在数据组织中各层数据的映射方式为：

对于第一层数据，分别用圆圈的大小和其填充颜色来可视化。其中，每个圆圈代表一个版块；圆圈半径表示对应在这个版块中帖子数目的多少，半径越大，帖子数目越多；圆圈填充的颜色：其中一个圆圈内填充的颜色与其他的圆圈不相同，表明下一层可视化显示了该板块中各个发帖区域帖子的统计信息。

对于第二层数据，每个圆圈代表一段时间内发布的帖子。与第一层相类似，圆圈半径对应在这个发帖时间段内帖子数目的多少；圆圈填充的颜色：表明下一层可视化显示了该发帖时间段内各个区域所发帖子的统计信息。第三层数据与上面两层相似。

第四层中，每个帖子对应一个小圆圈，分别用圆圈半径大小及其距离中心点的半径距离来可视化回复关系。圆圈半径：对应帖子的回复的人次，回复量越大，其半径也越大。圆圈与中心的距离：对应帖子发表的时间，离中心越近，发布时间靠近月初，离中心越远，发帖时间越靠近月末。

4 海量数据的动态可视化

对于多维海量的论坛数据，用一种可视化方式无法全面揭示出其规律。这里针对不同方面的数据集，提出了 3 种动态可视化方式：平行坐标、层次关系图以及参与者回复曲面关系图来表现数据集的特点，以使从不同角度可视化论坛数据。

4.1 基于平行坐标的论坛热点分析

平行坐标是一种可视化多维数据的可视化方式，其将 n

维数据(属性空间)通过 n 条等距离的平行轴映射到二维平面上,每一条轴线代表一个属性维,轴线上的取值范围从对应属性的最小值到最大值均匀分布。这样,每一条数据记录的各个属性值均可在相应的属性轴上标记出一点,这些点连成的折线即表示该条数据记录,多条折线可表示多条记录。它能使用户直观地看到数据集的全貌,分析各对象同一属性值的分布、分析各属性之间的关系,还可进行聚类分析等,以找出数据中隐含的规律。

在网络论坛数据的可视化中,单用层次的可视化方式,能表现数据一方面的信息,对于信息维数很多、数据之间关系复杂的网络论坛数据来说,仅仅用层次化的可视化方式还不够,不能展现全部的规律。具体地,针对海量的论坛信息,要找到某一阶段的热点信息是非常难的。这里采用平行坐标对论坛热点的转换进行可视化,展现论坛数据在某一段时间内的特点。

平行坐标的轴分别表示帖子发起的时间、热门程度、所属版块、持续时间、帖子内容长短以及发帖人 id 等之间的关系,如图 2 所示,绿线是某一个热点问题,从图中可以非常清晰地看到在不同的时期该问题的关注程度。

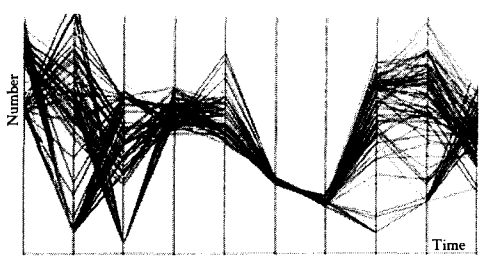


图 2 论坛数据的平行坐标可视化

4.2 参与者间回复曲面关系图

层次化的可视化方式以及平行坐标主要针对论坛中的帖子信息,但是无法反映作者的参与。我们进一步选取论坛中参与者的回复关系,来可视化不同人的参与情况。

由于论坛参与者之间是一个多对多的复杂网络关系。我们将每个参与者的 id 号抽取出来,组成一个参与者回复关系的网络图。但是如何表现这种复杂的论坛回帖关系?我们提出了一种曲面关系图的形式来反映以某个参与者为中心的发帖回贴关系。

具体地,每一个圆圈都代表一个作者 id,圆圈的大小、颜色以及圆圈之间的连线等都被赋予某些可视化信息。圆圈大小对应参与者 id 积分,与参与者 id 积分成正比。圆圈颜色对应于参与者发布的帖子被其他参与者回复的多少,越靠近红色表明被回复的次数越多,越靠近蓝色表明被回复的次数越少。绿色圆环的中心圆圈是可视化的主要参与者。其他圆圈与中心圆圈的连线:这些圆圈所代表的作者 id 回复过中心圆圈所代表的参与者 id。绿色圆环内的圆圈:中心圆圈代表的参与者 id 回复过绿色圆环内的圆圈所代表的参与者 id,并且回复次数越多,距离中心圆圈越近。绿色圆环外的圆圈:表示中心圆圈代表的参与者 id 号没有回复过绿色圆环外圆圈所代表的参与者 id,但是绿色圆环外的圆圈所代表的参与者 id 回复过中心圆圈代表的 id 发布的帖子,绿色圆环外的圆圈位置随机排布。对中心圆圈所代表的参与者 id 号没有回复的作者 id 没有在图中给出。

如图 3 所示,左图的中心为某一个作者 id,可以清晰地看到不同作者之间的关系,右图中我们将曲线关系图扩展成三维,以球形来展现不同回帖之间的关系,表现的空间和数量更大一些。

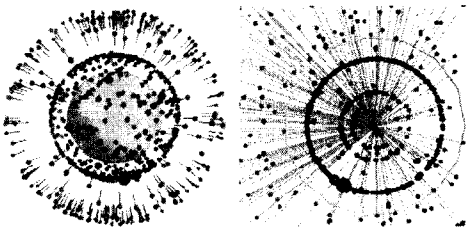


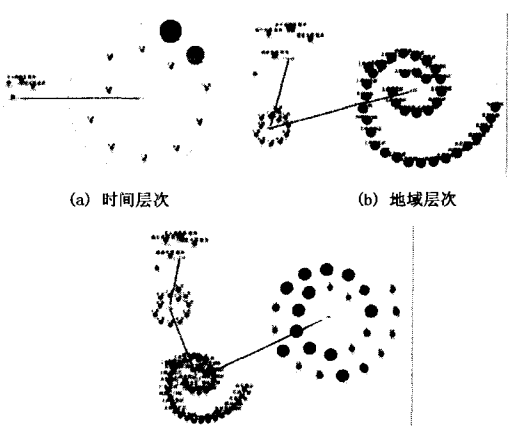
图 3 回复关系曲面图

4.3 论坛信息的层次交互可视化

为了更好地进行可视化分析,交互技术是必不可少的。我们设置了方便的交互方式,可以在不同层次之间切换,并可以对任意层次的数据集显示图进行放大、缩小、平移和旋转操作,从而可以方便地查看可视结果;可以支持海量的数据信息交互显示。

采用上面的方法,我们对篱笆论坛(http://bbs.sh.libaclub.com/)的数据进行了可视化分析。硬件平台是 Intel(R) Pentium(R) 3.4 GHz CPU,2 GB 内存的微机,这里我们收集了从 2008 年到 2010 年一年多的数据,对其进行了交互可视化。

图 4 显示了层次可视化方法的交互过程,初始时选中了"甜蜜小屋"版块,这是第一层的可视化结果,图 4(a)是选中"12 月"后显示以上版块中的不同时间分布;然后可以选择不同地域下的分布,这里的地域是发帖作者的地域分布属性,图 4(b)为选中了"上海市浦东区"这一版块;最后一层是不同主题贴的信息。



(a) 时间层次 (b) 地域层次

(c) 回帖主题层次

图 4 层次可视化

我们采用层次的方法来进行可视化,算法效率和支持的数量得到了大大提高,表 1 为我们的方法与相关方法的对比。

表 1 本文算法与前人方法的对比

方法	可支持的记录数据量	算法效率
本文的层次可视化方法	>=100000 条	1 秒
本文的平行坐标方法	10000 条	0.1 秒
Fisher 等人的方法[4]	2000 条	3 秒
Heer 等人的方法[6]	500 条	—

结束语 本文针对网络论坛数据的特点,建立了论坛数据集的层次关系和可视化映射过程,然后采用平行坐标、曲面

关系图、层次交互图3种方式分别展示了论坛中的不同属性数据，并对结果进行了总结和分析，结果表明，其能够方便地进行海量论坛数据的层次多维可视化。

网络论坛数据的可视化为管理人们在论坛中的行为模式提供了一种直观的方法，对论坛的管理人员和其他分析人员有一定的指导意义。由于网络论坛数据的复杂性和大规模性，海量数据的可视化分析速度相对较慢，如何加快处理的速度是我们下一步的工作，以及进一步在网络论坛数据中挖掘出更多的聚类信息，如社会群体事件的信息。

参 考 文 献

[1] Smith M A, Fiore A T. Visualization Components for Persistent Conversations[C]// Proceedings of the SIGCHI conference on Human factors in computing systems. New York, NY, USA: ACM, 2001: 136-143

[2] Viégas F B, Smith M. Newsgroup, Crowds and AuthorLines: Visualizing the Activity of Individuals in Conversational Cyberspaces[C]// Proceedings of the 37th Hawaii International Conference. Waikoloa, Big Island, Hawaii, 2004

[3] Welser H, Gleave E, Fisher D, et al. Visualizing the Signatures of Social Roles in Online Discussion Groups [J]. Journal of Social Structure, 2007, 8(2): 1-13

[4] Fisher D, Smith M, Welser H T. You are who you talk to: Detecting roles in Usenet newsgroups [C]// Proceedings of the 39th Hawaii International Conference. Waikoloa, Big Island, Hawaii, 2006

[5] Paolillo J C. Visualizing Usenet: A Factor-Analytic Approach [C]// Proceedings of the 33rd Hawaii International Conference. Waikoloa, Big Island, Hawaii, 2000

[6] Heer J, Boyd D. Vizster: Visualizing Online Social Networks[C]// Proceedings of the 2005 IEEE Symposium on Information Visualization. Washington, DC, USA: IEEE Computer Society, 2005: 33-40

[7] Viegas F, Golder S, Donath J. Visualizing email content: Portraying relationships from conversational histories[C]// Proceedings of the SIGCHI conference on Human Factors in computing systems. New York, USA: ACM, 2006: 978-988

[8] Kerr, Bernard. Thread Arcs: An Email Thread Visualization[C]// Proceedings of the 2003 IEEE Symposium on Information Visualization. Washington, DC, USA: IEEE Computer Society, 2003: 211-218

[9] Rohrer R, Swing E. Web-based information visualization [J]. IEEE Computer Graphics and Applications, 1997, 7(8): 52-59

[10] Rohall S L, Gruen D, Moody P, et al. ReMail: A Reinvented Email Prototype[C]// Proceedings of ACM Human Factors in Computing Systems. 2004: 791-792

[11] Venolia G, Neustaedter C. Understanding Sequence and Reply Relationships within Email Conversations: A Mixed-Model Visualization[C]// Proceedings of ACM Human Factors in Computing Systems. 2003: 361-368

[12] 裴新. 网络中极大 k-plex 发现算法和网络社群简历挖掘研究[D]. 北京：北京邮电大学，2008

[13] 黄雄伟，陈定方，祖巧红. Web 数据挖掘可视化研究与应用[J]. 湖北工业大学学报，2009，24(4)：54-56

[14] 王柏，吴巍，徐超群，等. 复杂网络可视化研究综述[J]. 计算机科学，2007，34(4)：17-23

[15] 蔡磊，龚健雅. 分布式海量多源空间数据的组织与网络可视化[J]. 测绘信息与工程，2009，34(6)：28-30

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

(上接第 194 页)

纳与总结方法的问题，本文提出了一种面向研讨环境的摘要生成方法。该方法首先采用概率混合模型对专家发言进行分析，抽取专家发言的话题集，再判断专家发言话题是否发生转变，一旦专家发言的话题转变，就生成相应的摘要，提供给专家，专家可对其进行提炼以及升华，为最终形成有效的决策方案提供信息支持。实验结果表明，本文提出的面向研讨环境的摘要生成方法合理、有效。

参 考 文 献

[1] 钱学森，于景元，戴汝为. 一个科学新领域—开放的复杂巨系统及其方法论[J]. 自然杂志，1990，13(1)：3-10

[2] 王寿云，于景元，戴汝为，等. 开放的复杂巨系统[M]. 浙江：科学技术出版社，1996：278-282

[3] 戴汝为. 系统学与中医药创新发展[M]. 北京：科学出版社，2008

[4] 李耀东，崔霞，戴汝为. 综合集成研讨厅的理论框架、设计与实现[J]. 复杂系统与复杂性科学，2004，1(1)：27-32

[5] 赵明昌，李耀东. 一个新的综合集成研讨厅软件框架[J]. 计算机工程与应用，2008，44(11)：1-4

[6] 戴汝为. 人-机结合的智能工程系统—— 处理开放的复杂巨系统的可操作平台[J]. 模式识别与人工智能，2004，17(3)：257-261

[7] 戴汝为，李耀东. 基于综合集成的研讨厅体系与系统复杂性[J]. 复杂系统与复杂性科学，2004，1(4)：1-24

[8] 操龙兵，戴汝为. 综合集成与决策[J]. 计算机研究与发展，2003，40(4)：531-537

[9] 刘春梅，戴汝为. 综合集成研讨厅专家群体评估结果的可视化[J]. 模式识别与人工智能，2005，18(1)：6-11

[10] 王丹力，戴汝为. 群体一致性及其在研讨厅的应用[J]. 系统工程与电子技术，2001，23(7)：33-37

[11] 王丹力，戴汝为. 专家群体思维收敛的研究[J]. 管理科学学报，2002，5(2)：1-5

[12] Bhandari H, Ito T, Shimbo M, et al. Generic Text Summarization Using Probabilistic Latent Semantic Indexing [C]// Proceedings of IJCNLP. 2008

[13] Hofmann T. Probabilistic Latent Semantic Indexing [C]// Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. 1999: 50-57

[14] Gong Y, Liu X. Generic text Summarization using relevance measure and latent semantic analysis [C]// Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001: 19-25

[15] Erkan G, Radev D R. LexPageRank: Prestige in Multi-Document Text Summarization [C]// Proceedings of EMNLP. 2004

[16] Mei Qiao-zhu, Zhai Cheng-xiang. Discovering Evolutionary Theme Patterns from Text-An Exploration of Temporal Text Mining [C]// Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining

[17] Zhai Cheng-xiang, Velivelli A, Yu Bei. A Cross-Collection Mixture Model for Comparative Text Mining[C]// Proceedings of KDD '04

[18] Cover T M, Thomas J A. Elements of Information theory [M]. Wiley, 1991