基于形态平行坐标系的层次数据可视化研究。

成生辉! 蒋志方1+,2, 孟祥旭1,2, 杨维强3, 辛若波2, 李宁2

1(山东大学 软件学院软件工程系,济南 250101)

2(山东大学 计算机科学与技术学院,济南 250101)

3(山东大学 数学院,济南 250100)

Research On Hierarchical Data Visualization Based On the Shape Parallel Coordinates System*

CHENG Sheng-Hui¹, JIANG Zhi-Fang^{1+,2}, MENG Xiang-Xu^{1,2}, YANG Wei-Qiang³, Xin Ruo-bo², Li-Ning²

¹(School of Software, Shandong University, Jinan 250101, China)

Abstract: This paper presents a method of Shape Parallel Coordinates System for data visualization. First, the axis is defined in the plane and the set of the axes is called the main shape. The lines set used to constraint and cooperate with the main shape is defined as the vice shape. The coordinate system formed by using the auxiliary shape set to set axes layout is defined as Shape Parallel Coordinates System, which has good plane figures simulation and superposition. According to the data dimension and structure, we give the Shape Parallel Coordinates System generation algorithm (including the single Shape Parallel Coordinates System generation algorithm and Shape Parallel Coordinates System superposition algorithm) and Transform Data algorithm. Finally, by means of L-system, we achieve a kind of hierarchical data visualization based on the Tree Shape Parallel Coordinates System.

Key words: Shape; Layout; Parallel Coordinates; Hierarchical Data; Visualization; L-system; Tree

摘 要: 本文提出了一种形态平行坐标系数据可视化方法。首先,在平面上定义了坐标轴,其集合称为主形态;用于约束和配合主形态布局的线集合定义为副形态。通过副形态约束和配合主形态布局而形成的坐标系定义为形态平行坐标系。形态平行坐标系具有良好的平面图形模拟性和可叠加性。根据数据的维数和结构,给出了形态平行坐标系的生成算法(包括单一形态平行坐标系生成算法和形态平行坐标系叠加算法)和数据映射算法。最后,利用 L-系统、实现了一种基于树状形态平行坐标系的层次数据的可视化。

关键词:形态:布局:平行坐标:层次数据:可视化:L-系统:树

中图法分类号: TP301 文献标识码: A

* Supported by the National High-Tech Research and Development Plan of China under Grant Nos.2009AA01Z304 (国家高技术研究发展计划(863)); Shandong Province Scientific and Technological Project, under Grant Nos.2009GG20001025(山东省科技攻关项目),Jinan Science and Technology Development Plan Project under Grant Nos.JNKF2010-00101-05 (科技发展计划项目)

² (School of Computer Science & Technology, Shandong University, Jinan 250101, China)

³(School of Mathematics, Shandong University, Jinan 250100, China)

⁺ Corresponding author: Phn: +86-531-88392498, Fax: +86-531-88392498, E-mail: zfjiang@sdu.edu.cn

作者简介:成生辉(1990-),男,山东省东营人,主要研究领域为数据与信息可视化,可视分析;**落志方**(1961-),男,博士,副教授,主要研究领域为人机交互与虚拟现实,数据与信息可视化,可视分析,数据挖掘;**孟祥旭**(1942-),男,博士,教授,博士生导师,主要研究领域为人机交互与虚拟现实,CAD/CAM,科学可视化;杨维强(1976-),男,博士,讲师,主要研究领域为金融数学。

1 引言

随着信息技术尤其是互联网和物联网技术的高速发展,传感器技术和数据采集技术及传输和存储技术的不断进步,各种类型的数据正呈现几何级数的增长,形成了各种类型的海量数据。这些数据维数呈现高维化,数据结构呈现复杂化。数据可视化[1]旨在借助于图形化手段,清晰有效地传达与沟通信息,便于人们发现隐藏在数据中的现象、知识和规律,挖掘更有价值的信息。

平行坐标已经成为信息可视化、可视化数据挖掘等领域的主流技术之一。平行坐标[2]是一种表达多维空间中点的一种几何投影方法,它把多维空间的数据点映射到二维空间,具有很好的数学基础和广泛的应用领域。通常,平行坐标以一组平行等距的水平或垂直坐标轴为基本坐标轴 $\{X_1,X_2,...,X_n\}$, n 维数据集 $E=\{e_{i,e_2},...,e_n\}$ 中的每一维属性对应一个坐标轴,其中 e_i 为列向量 $e_i=(v_{il,v_i},v_{i2},...,v_{ij},...)^T$ 。每一个坐标轴上连续或离散的值表示每一维的属性值。一个 n 维数据可用平行坐标上的一条折线表示,顶点在坐标轴上的取值即为相应数据的属性取值。

Alfred Inselberg[3]从射影几何和对偶特性的角度阐述了平行坐标的数学基础,并将其应用于计算几何。Wegman[4]将平行坐标应用于统计领域中的多元数据分析。Rosario 等[5]采用 Multiple Correspondence Analysis 的方法将类别转换为数字,使得平行坐标不仅可以显示连续变量[6],而且可以显示类别变量,从而使得每个轴的空间能被更有效地使用。K.T. McDonnell 等 [7]提出了 Illustrative Parallel Coordinates 的方法,采用密度图代替原始数据显示来充实和完善平行坐标[8]。Y.-H. FUA 等[9]研究了基于分层聚类的多分辨率视图方法,利用平行坐标上的变化来表达聚合的信息。H. Zhou 等[10]给出了在平行坐标中基于几何视觉的隐式提升聚类效果的方法。袁晓如等[11]结合散点图和平行坐标的优点,提出了在平行坐标里面引入散点图的方法。Julian Heinrich 等[11] 为实现空间连续数据的可视化,采用连续散点图的概念推导出了平行坐标的密度模型,提出了连续平行坐标的方法。Jarry H.T. Claessen 等[13]提出了灵活的连接的坐标轴方法,可以让用户自定义坐标轴的位置。

平行坐标的优点是表达多维数据非常直观,而不必使用矢量或其它可视图标。缺点是表达维数决定于屏幕的水平宽度,当维数增加,引起垂直轴靠近,辨认数据的结构和关系稍显困难;坐标间的依赖关系很强,垂直平行轴之间的排列顺序也是影响发现数据之间关系的重要因素[14];对于具有复杂结构的数据,平行坐标系不能很好地表现数据的结构特性。

层次数据是生活工作中常见的数据类型,经典的层次数据可视化技术[15]有 Robertson[16]等提出的 Cone tree,Shneiderman [17]等提出的 Tree-map[18],Lamping[19]等提出的 Focus+Context 技术 Hyperbolic tree。但是这些方法没有充分挖掘层次数据本身的形态特性。事实上,层次数据通常用树结构存储和表示,因此也可以采用树的生物形态对层次数据进行可视化。L-系统是有关生长发展中的细胞交互作用的数学模型[20] ,被广泛应用于植物生长过程的研究和模拟,可以用该模型来生成各种形态的树。L-系统可定义为一个四元组 $G = \{V,S,\omega,P\}$,其中 V为变量符号集合,S为常量符号集合,M为初始状态串,M为产生式规则[21]。自初始状态开始迭代套入 L-系统的产生式规则,通过龟形命令[22],可以生成相应的图形。

本文分析了层次数据的结构特点,研究了平行坐标数据可视化方法,给出了坐标轴和用以辅助坐标轴布局形态的副形态的定义,并以此为基础提出了形态平行坐标系。利用 L-系统,实现了基于一种树状形态平行坐标系的时序型层次数据的可视化。

2 形态平行坐标系

2.1 形态平行坐标系定义

2.1.1 坐标轴

平面上,一条规定了原点、正方向和单位长度的线称为坐标轴,记为 \mathbf{X} ,表示为

$$X = \{l, O, D, U\}$$

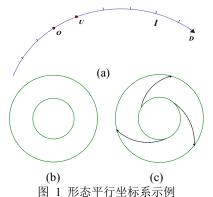
其中l为坐标轴线,O为坐标轴的原点,D为l 的正方向,U 为单位长度。如图 1 (a)。

2.1.2 主形态

平面上,由坐标轴构成的集合称为主形态,记为 M, $M = \{X_1, X_2, ..., X_n\}$ 。

其中n为坐标轴的总数。

2.1.3 副形态



平面上,用于约束和配合主形态布局的线集合称为副形态,记为 V。如图 1(b)。

2.1.4 形态平行坐标系

平面上,通过副形态约束和配合主形态布局而形成的坐标系称为形态平行坐标系,记为 SPCS。形态平行坐标系是由坐标轴集和副形态的并集构成的线集合,即 SPCS = $M \cup V$ 。当 M 和 V 可以构成形态坐标系时,SPCS 可表示为 SPCS = $M \mid V$ 。图 1(c)为图 1(a)构成的坐标轴集和图 1(b)并后的形态平行坐标系。

2.2 形态平行坐标系的性质

性质 1: 平面图形可模拟性

形态平行坐标系可以是平面上任意由线构成的图。

证明:设 G 是平面上由线构成的图,显然 G 是平面上的线集。由定义 2.1.4 可知,形态平行坐标系也是平面上的线集。故存在形态平行坐标系 SPCS,使得 SPCS=G。

SPCS 的构造方法: 取任意集合 $A \subseteq G$,令 M=A,则 V=G-M。由定义可知,可以得到 SPCS 为 $\langle G-A|A \rangle$ 。如图 2 所示。

图 2(a)为模拟饼图的形态平行坐标系,其中 M 为外面的圆,V 为两个扇形的分界线集合。图 2(b)为模拟星型图的形态平行坐标系,其中 M 为外面的轮廓线,V 为轮廓与中心相连线的集合。

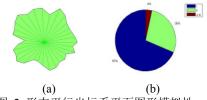


图 2 形态平行坐标系平面图形模拟性

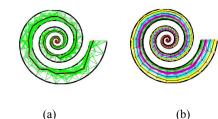


图 3 形态平行坐标系的可叠加性

性质 2: 可叠加性

对于任意 n 个形态平行坐标系 $SPCS_1$, $SPCS_2$,..., $SPCS_n$, 其副形态分别记为 V_1, V_2 ,..., V_n , 主形态分别记为 M_1, M_2 ,..., M_n , 令 $V = V_1 \cup V_2 \cup ... \cup V_n$, $M = M_1 \cup M_2 \cup ... \cup M_n$, 则新的形态平行坐标系可由多个形态坐标系叠加而成,即 $SPCS = SPCS_1 \cup SPCS_2 \cup ... \cup SPCS_n = \langle M | V \rangle$ 。

证明:由于 $M_1, M_2, ..., M_n$, $V_1, V_2, ..., V_n$ 均为线集,则 M 和 V 仍为线集,故 M 和 V 可以分别作为主形态和 副形态,由定义可知 SPCS 为形态平行坐标系,即 SPCS = $\langle M|V\rangle$ 。如图(3)所示。图 3(a)为一个模拟螺旋线 的形态平行坐标系,V 为外轮廓线和内轮廓线,M 为同一角度两轮廓线相夹的线集合。图 3(b)为 4 个模拟螺旋线的形态平行坐标系叠加后生成的新坐标系。

2.3 形态平行坐标系中的算法

算法 1: 单一形态平行坐标系生成算法 Generate_Single_SPCS(E) 1° 根据数据的维数,设计坐标轴个数 n_{\circ} 2° 根据数据的结构,设计 X 和 V_{\circ} 3° 根据形态平行坐标系的定义,生成 $SPCS_{\circ}$ return $SPCS_{\circ}$

算法 3: 形态平行坐标系生成算法
Generate_SPCS(E)
listE=null
listC=null
flag=false
if E要分割
do flag= false
每部分记为 Ei, listE.add(Ei)
for each Ei in listE
do listC.add(Generate_SPCS(Ei))
if flag=false
then return Generate_Single_SPCS(E)
return Superposition_SPCS(listC)

算法 2: 坐标系叠加算法 Superposition_SPCS(SPCSset) $M=\emptyset$, $V=\emptyset$ for each SPCS $_i$ in SPCSset do $M=M\cup get(M_i)$ $V=V\cup get(V_i)$ SPCS = (M|V)

return SPCS

算法 4: 数据映射算法 Transform_Data(E,SPCS)

 $1^{\circ}e_{i}$ 中的数据在 SPCS 中的坐标轴 X_{i} 上表示。 v_{ij} 的位置是在 l_{i} 上沿着正方向 D 与 O_{i} 距离为 $v_{ij}U_{i}$ ($v_{ij}U_{i}$ 表示两个它们的乘积)所对应的点,记为 $P_{v_{ij}}$,表示为 $P_{v_{ij}}=O_{i}+v_{ij}U_{i}$ 。

 2° 不同坐标轴上的 $P_{\nu_{y}}$ 可以通过连线、颜色填充等渲染方式展现数据。

3 基于形态平行坐标系的层次数据可视化

3.1 数据来源

本文所要进行可视化的数据是从 2002 年 1 月到 2011 年 12 月的某市每天的可吸入颗粒物(PM10),二氧化硫(SO2)和二氧化氮(NO2)监测数据。该数据具有层次结构(年 \rightarrow 月 \rightarrow 日)。通过这些实际监测的污染物浓度,可以分别计算得到各种污染物的得分值,分别记为 $API_{PM_{10}}API_{SO2}$, API_{NO2} 。空气污染指数(Air Pollution Index,简称 API)的计算方法[26]如式(1)。根据上述数据,实现对 API 的可视化。

$$API = \max(API_{SO2}, API_{NO2}, API_{PM_{10}})$$
 (1)

得分值与 API 值相等的污染物称为首要污染物。

3.2 形态平行坐标系结构分析

由于 API 与 PM10, SO2, NO2 有关,为了生成 API 坐标系,可以分别生成 PM10, SO2, NO2 坐标系,然后将其叠加。由于 PM10, SO2, NO2 具有层次结构,可以用树状形态实现可视化。API 坐标系为森林。

首先将 API 坐标系 C_{API} 分为 C_{PMI0} , C_{SO_2} 和 C_{NO_2} 。将 C_{PMI0} , C_{SO_2} 和 C_{NO_2} 又分为 C_{limb} , C_{leaf} 和 C_{pie} 。 其中 C_{limb} 又分为 C_{trunk} 和 C_{branch} , C_{leaf} 分为 $C_{redleaf}$ 和 $C_{greenleaf}$ 。 C_i (i=PM10,SO2,NO2)表示该市污染物 i 的坐标系,以 树状形态展现。 C_{trunk} 为树干坐标系,自下而上依次表示 2002-2011 年,每个树干展示该年每天的监测数据。 C_{branch} 为枝坐标系,表示 1-12 月,每个枝展示该月每天的监测数据。 $C_{redleaf}$ 和 $C_{greenleaf}$ 分别为红叶坐标系和绿叶坐标系,叶脉表示 1-28\29\30\31 天,从水平向右沿逆时针方向依次生成。 $C_{redleaf}$ 每个叶脉的长度为该天的污染物得分值, $C_{greenleaf}$ 每个叶脉的长度为该天 API。 C_{pie} 为饼图坐标系,表示 2002-2011 年,每个饼图表示该污染物在该年中的等级分类,其半径为该污染物在该年份作为首要污染物的次数。

3.3 基于L-系统的绘制形态平行坐标算法

根据算法 1, 2, 3 用 L-系统绘制形态平行坐标的步骤如下。本文中, L-系统中每个参数的意义如表 1 所示。

1°执行如下 L-系统,产生字符串。 V:{Init,Tree,Root,Internode,Branch} S:{F,+,-,[.],O_i,T_i} 初始状本。R

初始状态: B 产生规则:

 $B \rightarrow Init \ (Tree)^{n_{tree}}$

Init → 读入 initpoint, θ_{init} , θ_{i} , n_{tree} , n_{branch} , n_{shape} , $n_{subranch}$, d_{root} , d_{branch} , $d_{internode}$, S_{branch} \circ

 $Tree \rightarrow +(\theta_{init}) Root ([Internode M Branch])^{n_{branch}}$

 $Root \rightarrow F(d_{root})$, i = 0, directflag = 1

 $Internode \rightarrow F(d_{internode})$

 $Branch \rightarrow O_i F(d_{branch}) T_i (S(S_{branch}))^{n_{shape}} (T')^{n_{subranch}}, i = i + 1$

M: $flag = I \rightarrow +(\theta)$, flag = I;

 $flag = 1 \rightarrow -(\theta), flag = 0$

2°用龟形命令生成图形。

表1 L-系统参数表

| 符号 | initpoint | $	heta_{	ext{init}}$ | θ | n_{tree} | n_{branch} | n _{shape} | n _{subranch} | d_{root} | d_{branch} | $d_{internode}$ | S_{branch} |
|--------------------|-----------|----------------------|----------|------------|--------------|--------------------|-----------------------|------------|--------------|-----------------|------------------------------------|
| 含义 | 初始点 | 初始方向 | 角度 | 树数 | 枝数 | 形状数 | 分支数 | 根长 | 枝长 | 节长 | 枝上形状 |
| value ₁ | (0,0) | π /2 | π /3 | 3 | 10 | 0 | 12 | 0.1 | 0.02 | 0.06 | 无 |
| value ₂ | 当前点 | 当前方向 | π/3 | 1 | 10 | 2 | 0 | 0.1 | 0.02 | 0.06 | C _{leaf} C _{pie} |
| value ₃ | (0.5,0) | π /2 | π /3 | 1 | 10 | 0 | 12 | 0.1 | 0.02 | 0.06 | 无 |
| value ₄ | (1,0) | π /2 | π /3 | 1 | 10 | 0 | 12 | 0.1 | 0.02 | 0.06 | 无 |

3.4 形态平行坐标生成

3.4.1 单一坐标系生成

根据形态平行坐标系的平面图形可模拟性和算法 1,4,交互地设计局部坐标系。 C_{pie} 见图 2(a), $C_{greenleaf}$ 如图 2(b) 所示, $C_{redleaf}$ 仿照 $C_{greenleaf}$,数据换为污染物的得分值。用 L- 系 统 绘 制 C_{trunk} 和 C_{branch} ,见 3.3 ,其 中, $M = \{X_i | X_i$ 为线段 $O_i T_i\}$, $V = \{$ 除去A的线段 $\}$ 。如图 4。图 4(a)中每个树枝代表一年,奇数和偶数年分开,每年

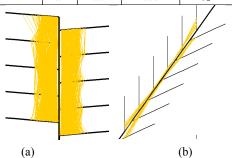


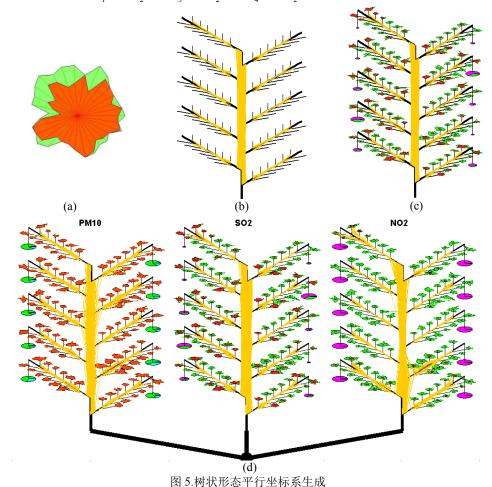
图 4 L-系统模拟枝干

同一天的数据相连。图 5 (b) 中每个树枝代表一月, 奇数和偶数月分开, 每月同一天的数据相连。 3.4.2 叠加生成 API 坐标系

根据可叠加性,根据算法 1,将生成的单一形态平行坐标系按如下方法叠加,生成 API 坐标系。

- (1) $C_{leaf} = C_{redleaf} \cup C_{greenleaf}$ 如图 5 (a) 所示。
- (2) C_{limb} = C_{trunk} ∪ C_{branch} 如图 5 (b) 所示。

在 L-系统中输入表 1 中 value₁ value₂ value₃ value₂ value₄ value₂ 时,完成所有叠加,绘制成 API 坐标系。



3.5 形态平行坐标系中的交互式数 据显示

为了更好地观察形态平行坐标 系,可以将其放大旋转等。为了更容 易地找到首要污染物,将 C_{API} 中的 C_{PM10} , C_{SO_2} 和 C_{N0_2} 同时局部放大, 实现坐标轴的联动。如图 6 所示。

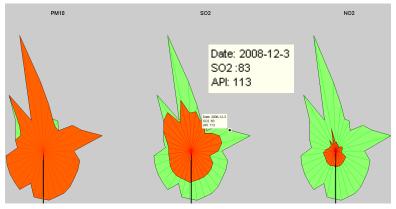


图 6 联动坐标轴

4 结束语

平行坐标是一种很好的数据可视化方法。随着数据的维数和结构复杂性的增长,对于数据的维数和结构的可视化,对于形态平行坐标系布局的研究将成为一个非常重要的问题。本文提出的形态平行坐标系数据可视化方法在这方面做了些工作,能够解决这部分问题。通过对某市 2002-2011 年的具有层次特点的 API 数据进行基于形态平行坐标系的可视化,取得了较好的整体和局部的可视化效果。实验结果表明,形态平行坐标系可以按照人们的意愿来自定义生成,有助于复杂结构的数据和信息的可视化,能帮助发现隐藏在数据中的知识。

References:

- [1] Yuan Xiaoru, Zhang Xin, Xiao He, Guo Hanqi Visualization Research Frontier and Prospect ISSN1674-9480 CN 11-5943/TP.Vol.2 No.4
- [2] 王绍敏, 孙晓静, 王克峰, 姚平经 应用平行坐标系进行可视化优化设计 计算机与应用化学 2004 年第 21 卷第 1 期
- [3] Alfred Inselberg. Parallel Coordinates visual Multidimensional Geometry and Its Applications 2009-08-28. ISBN: 0387215077.
- [4] Wegman E. Hyper dimensional data analysis using parallel coordinates[J]1990, 85 (411): 664—675.
- [5] Geraldine E. Rosario, Elke A. Rundensteiner, David C. Brown, Matthew O. Ward, and Shiping Huang. Mapping nominal values to numbers for effective visualization. In Proceedings of the IEEE Symposium on Information Visualization. IEEE Computer Society, 2003.
- [6] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In Proceedings of the 1st conference on Visualization '90, pages 361–378. IEEE Computer Society Press, 1990.
- [7] K. T. McDonnell K. Mueller. Illustrative parallel coordinates. Eurographics/IEEE-VGTC Symposium on Visualization. Vol 27(3), 2008.
- [8] J. J. Miller and E. J. Wegman. Construction of line densities for parallel coordinate plots. Springer-Verlag New York, Inc., 1991, pages 107 123.
- [9] Y.-H. Fua, M. O.Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In Proceedings of IEEE Visualization, pages 43 - 50, Oct. 1999.
- [10] H. Zhou, X.Yuan, H.Qu, W.Cui and B.Chen. Visual clustering in parallel coordinates. Computer Graphics Forum, 27(3):1047 -1054,2008.
- [11] Xiaoru Yuan, Peihong Guo, He Xiao, Hong Zhou, and Huamin Qu Scattering Points in Parallel Coordinates IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 15, NO. 6, NOVEMBER/DECEMBER 2009
- [12] Julian Heinrich and Daniel Weiskopf, Continuous Parallel Coordinates
- [13] Jarry H.T. Claessen and Jarke J. van Wijk Flexible Linked Axes for Multivariate Data Visualization
- [14] YU SHDONG ZHANGSHENGNAN Study on Enhance Visibility of Parallel Coordinates TP391.41 CNKI: ISSN: 1008-0570.0.2007-06-128
- [15] JING Min chang, SUN Jie li Study on Methods of Large Hierarchical Information Visualization Information Science Vol. 26, No. 4
- [16] George G.Robertson, Jock D.Mackinlay. and Stuart K.Card Cone Trees: Animated 3D Visualizations of Hierarchical Information [EBP OL] 2007 07 25.
- [17] Ben Shneiderman Discovering Business Intelligence Using Treemap Visualizations April 11, 2006
- [18] James Abello 1, Stephen G. Kobourov 2, and Roman Yusufov 2 Visualizing Large Graphs with Compound-Fisheye Views and Treemaps J. Pach (Ed.): GD 2004, LNCS 3383, pp. 431–441, 2004.
- [19] J. Lamping ,R. Rao. The hyperbolic browser: A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. [J]. Visual languages and Computeing ,1996 ,7(1):33 55.
- [20] Aristid Lindenmayer. Mathematical models for cellular interaction in development. [J]. Theoretical Biology, 18:280—315, 1968
- [21] Przemyslaw Prusinkiewicz, Aristid Lindenmayer. The Algorithmic Beauty of Plants, 1996. softcover, 240 pages, ISBN 0-387-94676-4.
- [22] Joanna L. Power A.J. Bernheim Brush Przemyslaw Prusinkiewicz David H. Salesin Interactive Arrangement of Botanical L-system Models ACM New York, NY, USA ©1999 table of contents ISBN:1-58113-082-1 doi>10.1145/300523.300548
- [23] Ambient air quality standard GB 3085-1996