第33卷 第17期 计 算 机 工 程 2007年9月
Vol.33 No.17 Computer Engineering September 2007
·网络与通信· 文章编号：1000—3428(2007)17—0140—03 文献标识码：A 中图分类号：O157

中国教育网的结构分析

张 宁
（上海理工大学管理学院，上海 200093）

摘 要：从静态结构特性、度相关性、网络的整体结构和网络联通集团的结构等方面对中国教育网的结构进行了研究。研究结果为充分了解中国教育网的结构特征提供了参考，为进一步提出好的搜索策略、改进搜索引擎的功能提供了思路。
关键词：万维网；拓扑结构；蝶形结图；连通集团

Analysis of China Education Network Structure

ZHANG Ning
(Business School, University of Shanghai for Science and Technology, Shanghai 200093)

【Abstract】 According to the topologic structure, correlation of degrees, bow tie picture, connected components' structure, this paper reaches the structure of China education network. The results help to understand the properties of the network and give some new ideas to improve searching strategy and searching engine function.
【Key words】 world wide Web; topology; bow tie picture; connected component

随着国际互联网的高速发展，万维网的规模在不断增加，为人们提供了便捷的信息交换和通信服务，成为各个领域的研究重点。对万维网的研究目前主要集中在拓扑结构[1~4]和演化模型[5~6]2个方面。了解万维网的结构及其演化规律对理论研究有许多实际意义[7]。对其拓扑结构理解得越深刻就越有可能设计出更好的搜索策略以进行分组和分类、改善浏览时间、改进搜索引擎的功能，还可提出更符合实际的模型来描述万维网的演变，从而产生新的演化算法思路。

中国教育网是万维网的一个子网，是由中国教科网中.edu.cn域名下的网页及其超链接构成的虚拟网络。在该网络中，所有网页都抽象为节点，而网页上由该网页指向其他网页的所有超链接都抽象为网络的有向边。由此构造了一个由366 422个节点和540 755条边构成的复杂有向网络[8]。

1 中国教育网的统计特性

由图论的知识可知，与一个节点关联的边数就是该节点的度。对于有向图，节点度有入度和出度之分，分别表示与该节点关联的入边数和出边数。根据统计研究，中国教育网节点出度概率$p_{out}(k)$和入度概率$p_{in}(k)$随节点度数k变化的双对数分布见图1和图2。

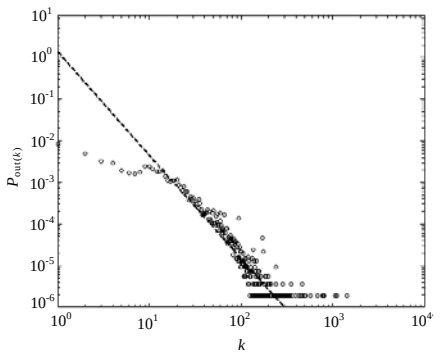


图1 节点出度分布

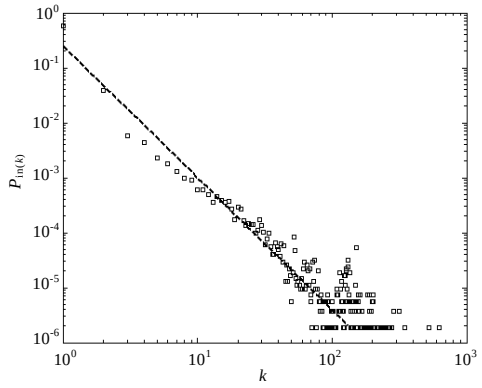


图2 节点入度分布

从图1和图2可以发现，中国教育网节点度分布图尾部呈幂律分布，$p_{out}(k)\sim k^{-r_{out}}$，其中，$r_{out}$=2.48；入度分布图基本呈幂律分布，尾部不是很平缓，$p_{in}(k)\sim k^{-r_{in}}$，其中，$r_{in}$=2.40。

中国教育网由各高校自建的网页所组成的子集构成，表1给出了中国教育网（.edu.cn）的统计特性[9,10]与Barabási等人研究网络（.nd.edu）的统计特性[11,12]相比较的结果。从表1中可以看出，中国教育网的规模比Barabási等人研究的网络规模略大，但它包含的边数远小于Barabási等人研究的网络，约是它的1/3，平均度也比后者小。

表1 网络统计特性比较

网络	节点数	边数	平均度	平均最短路径	入/出度指数	群聚系数
www.edu.cn	366 422	540 750	1.476	8.957	2.40/2.48	0.022 24
www.nd.edu	325 729	1 469 680	4.51	11.2	2.10/2.45	0.29

基金项目：国家自然科学基金资助项目(70571074/G0116)；上海市重点学科建设基金资助项目(T0502)；上海市自然科学基金资助项目(06ZR14144)
作者简介：张 宁(1956－)，女，副教授、硕士，主研方向：计算机网络与应用，系统分析与集成
收稿日期：2007-01-23 **E-mail：**zhangn@citiz.net

中国教育网的平均最短路径小于 Barabási 等人研究的网络，平均路径长度为 8.957，且路径长度的分布满足泊松分布，这说明虽然同是万维网中的教育类子网，二者的网络拓扑结构略有差异。在中国教育网中，许多网页都与全国各高校的主网页超链接，这可能是导致中国教育网平均路径较小的主要原因。

2 个网络的群聚系数用式(2)计算的结果见表 1，中国教育网的群聚系数小于Barabási等人研究的网络，但远大于相同规模的随机网络（其群聚系数小于 10^{-6}）。综上所述，中国教育网是一个出/入度服从幂律分布的无标度网络，具有小世界特性的有向复杂网络。

$$C_i = \frac{\text{包含节点}i\text{的三角形个数}}{\text{以节点}i\text{为中心的三点组个数}} \tag{1}$$

$$C = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{2}$$

2 中国教育网的度相关性分析

度相关性分析主要是研究中国教育网节点度与其邻接节点平均度的关系，假设节点 i 的度是一个随机变量 K，那么其邻接节点的平均度也是一个随机变量 $\overline{K}$，度相关性分析就是对 K 和 $\overline{K}$ 两个随机变量进行分析，计算出 K 与 $\overline{K}$ 的相关系数。相关系数的计算见式(3)。

$$\rho_{K\overline{K}} = \frac{E\{[K - E(K)][\overline{K} - E(\overline{K})]\}}{\sqrt{D(K)}\sqrt{D(\overline{K})}} \tag{3}$$

其中，$E(K) = \sum_{i=1}^{n} k_i p_i$ 为随机变量 K 的数学期望；$D(K) = \sum_{i=1}^{n} [k_i - E(k)]^2 p_i$ 为随机变量 K 的方差。

由于中国教育网是有向图，因此节点度的相关性分析包括几个方面，如节点 i 的入度与其邻接父（子）节点平均入度的相关性，节点 i 入度与其邻接父（子）节点的平均出度的相关性，节点 i 出度与其邻接父（子）节点入度的相关性，节点 i 出度与其邻接父（子）节点出度的相关性等一共 8 组相关性，中国教育网度相关系数的计算结果见表 2。

表 2　中国教育网节点度与其邻接节点平均度的关系

节点入度与其邻接父节点平均入度的相关系数	0.023
节点入度与其邻接父节点平均出度的相关系数	0.010
节点出度与其邻接父节点平均入度的相关系数	0.038
节点出度与其邻接父节点平均出度的相关系数	-0.018
节点入度与其邻接子节点平均入度的相关系数	0.041
节点入度与其邻接子节点平均出度的相关系数	0.094
节点出度与其邻接子节点平均入度的相关系数	0.116
节点出度与其邻接子节点平均出度的相关系数	0.208

从表 2 可以看出，中国教育网中节点度与其邻接节点平均度的相关性都很弱，节点度与其邻接父节点的平均度相关性很小。此外，节点的入度与其邻接子节点的度相关性也很小，只有节点的出度与其邻接子节点的出(入)度相关性相对较大，其中，节点出度与其邻接子节点的平均出度成最大的正相关。网络中的节点与其邻接子节点的度相关性都大于与其邻接父节点的度相关性，说明节点对于其邻接子节点度的影响大于对其邻接父节点度的影响。

表 2 中节点的出度与其邻接父节点的出度呈负相关，这一现象表明，对于一个承载一定内容的高校网站，从主页到各子主页，如果子页本身包含很多方面的内容，那么只需要划分个数较少的子主页就可以容纳网站的所有内容，即导致主页的出度较少；相反，如果各子页只包含几个方面的内容，那么在主页下就得划分更多的子主页来容纳其他内容，主页的出度就会相应增大，即子页的出度与主页的出度呈负相关。

3 中国教育网的蝶形结图

Broder等人用蝶形结图描述了万维网的宏观结构[4]，提出万维网主要由 5 个部分组成：核芯，入集，出集，枝蔓和其他部件，如图 3 所示。其中，核芯由网络中最大强连通分量构成；入集由能够连接到核芯的节点构成；出集由核芯中的节点所能到达的节点构成；枝蔓是指不包含在核芯中、由入集中可到达的节点和可到达出集中的节点构成；除此之外的剩余节点构成了其他部件。

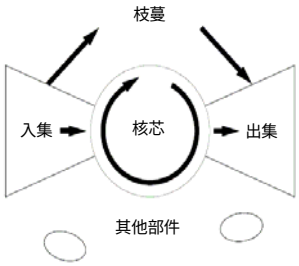


图 3　万维网的蝶形结图

通过对中国教育网进行分析，发现其最大的连通分支包含 315 484 个节点，占了整个网络的 86.1%。该网络的 366 422 个节点中，有 308 975 个节点均为入度为 1 而出度为 0 的节点，删除这些节点，就得到一个包含 57 447 个节点、231 775 条边的主子网，其平均度为 4.034。分别计算这个主子网的 5 个组成部分，并且把各组成部分的内部边按这 5 个部分进行分类，结果见表 3。

表 3　中国教育网主子网节点与边的分类及其所占比重

组成部分	核芯	入集	出集	枝蔓	其他
内部节点数/个	5 417	7 365	28 362	7 359	8 944
占节点总数比例/%	9	13	49	13	16
内部边数/条	37 788	11 570	32 933	9 770	36 557
占边总数比例/%	29	9	26	8	28

从表 3 可以看出，中国教育网中主子网的核芯节点数只占总节点数的 9%，而边却占总边数的 29%，超过其他任何组成部分的内部边数，使得核芯的平均度为 6.969，高于主子网的平均度 4.034，这说明核芯内部是强连通的。

除了各组成部分内部存在边连接以外，各组成部分之间也存在边连接，边分布情况的统计结果见表 4。

表 4　各组成部分之间的边连接

组成部分间的连接	入集到核芯	核芯到出集	入集到枝蔓	枝蔓到出集
边数	25 788	54 284	11 957	11 128
占边总数比例/%	25	53	11	11

从表 4 可以看出，从入集到核芯的边数以及从核芯到出集的边数之和占各组成部分之间连接边总数的 78%，这说明核芯和入集之间以及核芯和出集之间存在大量的边相连。

4 中国教育网的连通集团结构

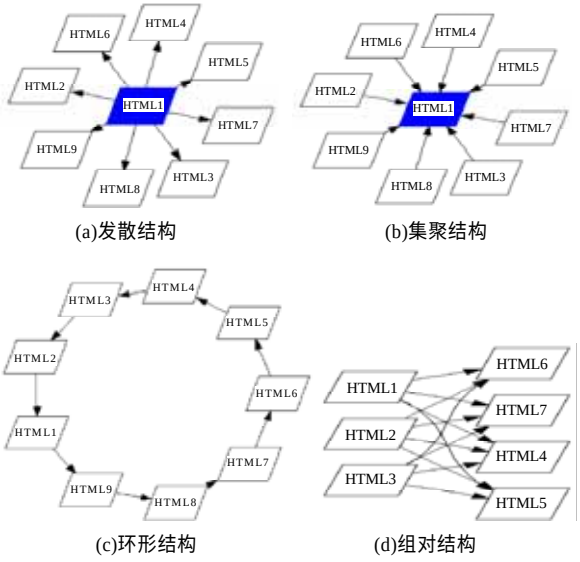
中国教育网中存在着一些规模不等的连通集团结构，这些结构内部高度连通，而各个集团之间只有较少的边相连接。从网络演化的角度来看，网络初始时存在一些小规模的连通集团结构，通过不断地加入新网页来丰富网站内容从而演化形成巨大的网络，这些小连通集团具有发散结构、集聚结构、环形结构和组对结构，这几种典型结构见图 4。

(1)发散结构中有一个中心节点，中心节点含有许多指向其他网页的超链接。中心节点包含某一相关主题，而中心节点指向的各个分支节点含有与该主题有关的一些内容。

(2)集聚结构也存在一个中心节点，许多网页都含有指向这一中心节点的超链接，这里的中心节点一般包含一些权威的或很有价值的内容，很多其他网页都含有指向它的超链接。

(3)环形结构中每一个节点都和相邻的节点首尾相连，形成环形，这种结构之所以大量出现是因为对于规模稍大的网站来说，从主页连接到各个终端网页的深度较大，为了便于浏览，在这种终端子网页中往往都会有返回主页等类似的超链接存在。

(4)组对结构由一些网页对{*i*,*j*}构成，它分为 2 部分：左边包含 *i* 个网页，右边包含 *j* 个网页，其中，左边的每一个网页都有指向右边网页的超链接，图 4(d)是 *i*=3，*j*=4 的情形。



(a)发散结构　(b)集聚结构

(c)环形结构　(d)组对结构

图 4　连通集团的结构

从图 4 可以看出，发散结构和集聚结构就是组对结构在 *i*=1，*j*=8 和 *i*=8，*j*=1 的特殊情况。在中国教育网中，这种组对结构大量存在，统计结果见图 5，说明中国教育网的演化是从组对结构开始通过不断增加新网页而逐渐发展壮大的。

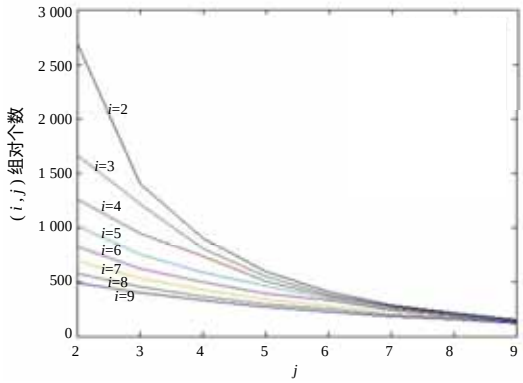


图 5　中国教育网的组对结构数量统计

5　结束语

中国教育网是一个具有小世界特性的无标度网络，网络中节点度与其邻接节点平均度的相关性都很弱，其中，只有节点的出度与其邻接父节点的平均出度呈负相关，其余都为正相关。该网络有一个主子网，主子网的节点数与边数分别占整个网络的 15.68%和 42.86%。主子网的蝶形结图中，核芯中的节点数仅占主子网总节点数的 9%，而核芯中的边数却占主子网总边数的 29%，核芯的平均度为 6.969，高于主子网的平均度 4.034 和整个网络的平均度 1.476，核芯内部是强连通的。网络中存在着一些规模不等的连通集团结构，这些结构内部高度连通，而各个集团之间则只有较少的边相连接，其中，组对结构大量存在。本文的研究结果有助于提出新的中国教育网演化思路和新的网络生成方法，弥补了文献[8]的演化算法费时的缺陷，也为进一步提出有效的搜索策略、改进搜索引擎的功能提供了思路。

参考文献

1 Albert R, Jeong H, Barabási A L. Diameter of the World Wide Web[J]. Nature, 1999, 401(9): 130-131.

2 Barabási A L, Albert R. Emergence of Scaling in Random Networks[J]. Science, 1999, 286(5439): 509-512.

3 Barabási A L, Albert R, Jeong H. Scale-free Characteristics of Random Networks: The Topology of the World Wide Web[J]. Physica A, 2000, 281: 69-77.

4 Broder A, Kumar R, Maghoul F, et al. Graph Structure in the Web[J]. Computer Networks, 2000, 33(1/6): 309-320.

5 Barabásii A L, Albert R, Jeong H. Mean-field Theory for Scale-free Random Networks[J]. Physica A , 1999, 272: 173-187.

6 Newman M E J, Strogatz S H, Watts D J. Random Graphs with Arbitrary Degree Distributions and Their Applications[J]. Phys. Rev. E, 2001, 64(2).

7 Donato D, Leonardi S. Mining the Inner Structure of the Web Graph[C]//Proc. of the 8th International Workshop on the Web and Database. 2005.

8 张　宁. 复杂网络实证研究——中国教育网[J]. 系统工程学报, 2006, 21(4): 337-340.

9 倪小军, 张　宁, 王美娟. 基于 MPI 的中国教育网最短路并行算法[J]. 计算机工程与应用, 2006, 42(12): 135-137.

10 Zhang N, Che H. The Topological Properties of China Education Network[C]//Proc. of KSS'06 Knowledge and Systems Sciences: Towards Knowledge Synthesis and Creation, Beijing. 2006.

11 Albert R, Barabási A L. Statistical Mechanics of Complex Networks[J]. Reviews of Modern Physics, 2002, 74(1): 47-97.

12 Newman M E J. The Structure and Function of Complex Networks[J]. SIAM Review, 2003, 45(2): 167-256.

（上接第 127 页）

图 8 是 $\bar{\gamma}$ 值的计算结果。根据上述讨论，$\bar{\gamma}$ 值则是越大效率越高，可扩展性越好。

4　结论

本文主要是从共享树 MPLS 组播的角度去研究 MPLS 组播，主要思路是如何构建二层的 MPLS 共享树和增强组播的可扩展性。本文提出的隧道和分枝节点相结合的方法，既可以实现共享树组播，也能增强可扩展性。

参考文献

1 Ooms D. Overview of IP Multicast in a Multi-protocol Label Switching (MPLS) Environment[S]. RFC3353, 2002.

2 Boudani A, Cousin B. A New Approach to Construct Multicast Trees [C]//Proc. of MPLS Networks the 7th International Symposium on Computers and Communications. 2002.

3 Boudani A, Cousin B, Jawhar C, et al. Multicast Routing Simulator over MPLS Networks[C]//Proceedings of the 36th Annual Simulation Symposium. 2003: 327-334.