第23卷 第2期 石家庄铁道大学学报(自然科学版) Vol. 23 No. 2
2010年8月 JOURNAL OF SHIJIAZHUANG TIEDAO UNIVERSITY (NATURAL SCIENCE) Aug. 2010

万维网的小世界效应探讨

赵正旭， 郭 阳， 刘贾贾， 龙 瑞

(石家庄铁道大学 信息科学与技术学院,河北 石家庄 050043)

摘要:在工程信息的长期安全保存和可靠调用过程中,存在着信息和其使用系统之间的兼容问题和对信息定期检查、移存、转化的时间和效率问题。目前国内外对这两类问题还没有可行的解决方法和技术。介绍了"六度分离"现象和经典的小世界网络模型,提出了WWW模型的架构,并在此基础上提出了万维网中任意两个网页间链接路径的搜索算法和计算万维网的特征路径长度,借此验证万维网具有小世界效应,为解决大量级和分散性信息的管理问题提供了可靠的科学依据。

关键词:六度分离;小世界效应;路径搜索;并行算法

中图分类号:TP393 **文献标识码**: A **文章编号**: 2095-0373(2010)02-0001-06

0 引言

匈牙利作家F Karinthy在1929年提出了"小世界现象"的论断。他认为,地球上的任何两个人都可以平均通过一条由6位联系人组成的链条而联系起来。在20世纪60年代,美国哈佛大学社会心理学教授斯坦利·米尔格兰姆(Stanley Milgram)通过设计一个连锁信件实验,提出了著名的"六度分隔"(Six Degrees of Separation)假说[1-2],大意为任何两个欲取得联系的陌生人之间最多只隔着6个人,便可完成两人之间的联系。当年,米尔格兰姆给内布拉斯加州奥马哈市随意选择的300多人发信,要求他们把他的这封信寄给波士顿市一个独一无二的"目标"人,分别由每个人独自联系。米尔格兰姆告诉每个发信人有关目标人的信息,包括姓名、所在地、职业,如果发信人不认识这个目标人,他们把这封信寄给他们认为有可能认识目标人的熟人。依此类推形成了发信人的链条,链上的每个成员都力图把这封信寄给他们的朋友、家庭成员、或同事熟人,以便使信件尽快到达目标人。米尔格兰姆发现,有60个链条最终到达目标人,链条中平均步骤大约为6,米尔格兰姆由此得出结论:任意两个人都可通过平均6个熟人联系起来。这就是六度分离理论的产生经过。

"六度分离"现象在学术上称为小世界效应(small world effect),小世界效应的定义是:若网络中任意两点间的平均距离L随网络格点数N的增加呈对数增长,即$L \sim \ln N$,且网络的局部结构上仍具有较明显的集团化特征,则称该网络具有小世界效应,这里的平均距离具有广泛的含义,例如在上述信件传递实验中,平均距离就是平均传递次数6。

"小世界效应"体现了一个似乎很普遍的客观规律:在如今的信息化时代,人们之间的关系已经完全社会化,任何两个素不相识的人都可能通过"六度分隔"产生联系。在看似庞大的网络中各元素之间的"距离"实际上是非常"近"的。有关统计表明,尽管万维网上信息数以亿计,但网络的特征路径长度L最多达到19[3],因此万维网是一个典型的小世界网络。本文旨在通过万维网的小世界模型结构探讨计算万维网特征路径长度的途径。

收稿日期:2010-04-13

作者简介:赵正旭,男,1960年出生,博士,教授。石家庄铁道大学信息科学与技术学院院长、东南大学长江学者特聘教授,精密仪器及机械专业博士生导师。曾任教于英国达毕大学计算机科学与技术系主任,终身教授,博士生导师,从事计算机科学与工程专业。主要研究方向为虚拟现实技术和应用。

基金项目:国家自然科学基金项目(60873208)

1 万维网拓扑模型

万维网作为当今人类社会信息化的标志,其规模正以指数速度高速增长。根据2005年中国互联网络信息资源数量调查报告结果显示,截至到2005年12月,全国网站数已经达到694 200个,网页总数约为24亿个,平均每个网站的网页数就有3 748个之多[4]。万维网作为一个典型的小世界网络,具有典型的小世界特性(较小的最短路径)和聚类特性(较大的聚类系数)。由此可以得出结论:任意两个网页之间是存在链接路径的,并且这种路径是可以被找到的。拟借助于小世界网络模型试图寻找网页之间的链接路径。

小世界网络模型(如图1所示)包括WS小世界网络模型[5]、NW小世界网络模型[6]以及一些其它的变形模型包括BW小世界网络模型等等[7]。其中Watts和Strogatz开创性的提出了小世界网络并给出了WS小世界网络模型;接着Newman和Watts又对WS小世界网络模型进行改进,提出了NW小世界网络模型,他们用随机化加边代替了随机化重连,从而避免了产生孤立节点的可能。因此WS和NW小世界网络模型是最为经典的小世界网络模型。

1998年,Watts和Strogatz提出了小世界网络的概念,并建立了WS模型。传统的规则最近邻耦合网络具有高聚类的特性,但并不具有小世界特性;而ER随机网络具有小世界特性但却没有高聚类特性。WS小世界网络模型就介于这两种网络之间,同时具有小世界特性和聚类特性,可以很好的来表示人工网络,比如万维网。

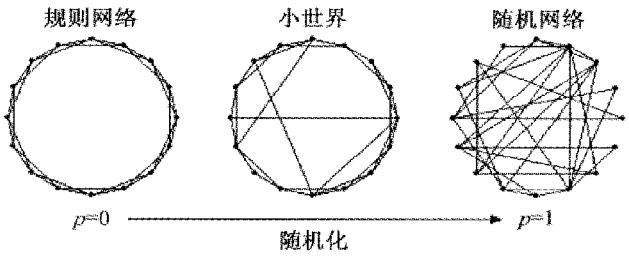

图1 三种基本网络模型

20世纪90年代,人们一直使用小世界网络构建WWW模型。后来,Barabasi和他的伙伴们通过分析网页上的链接信息后发现,绝大多数网页都拥有较少数目的链接,只有极少数的网页拥有较多的链接,比如一些门户网站。Barabasi认识到小世界网络模型并不适合WWW网络,因此提出了复杂网络的服从power-low度分布的Scale-free模型[8]。

由于万维网具有规模大、增长快的特点,很难把它作为实验的对象来进行研究和分析。对万维网所引发的一系列网络问题的研究往往只能基于网络拓扑模型进行,然而,由于万维网的结构变化很快,限制了对万维网拓扑结构的确定。虽然可以了解万维网的大致特征,但是无法构造出它的详细拓扑图。所以,万维网的拓扑模型只有根据已有的信息尽量准确地、有意义地反映真实的网络拓扑的实际情况。

2 网络统计量

网络统计量主要包括度和度分布、平均最短路径、聚类系数等等,这些统计量可以深刻的揭露网络的内部特性,下面先分别对这些统计量做简单的介绍。

(1)度与度分布。度(degree)是单独节点的属性中简单而又重要的概念。节点的度是指与该节点相关联的边的条数,也就是指与该节点连接的其它节点的数目。度分布(degree distribution)是指网络中各节点具有的度的分布。

(2)平均最短路径。网络的平均最短路径(average shortest paths)可以对网络的连通性进行较好地描述。网络中两个节点i和j之间的距离d定义为连接这两个节点的最短路径上的边数。网络的平均最短路径长度L定义为任意两点之间的最短路径的平均值。

(3)聚类系数。聚类系数(clustering coefficient)表征的是网络的聚类特性,也就是群落特性。一般假设网络中的节点i与k条边关联,即与另外k个节点相连。显然,在这k个节点之间最多可能有$k(k-1)/2$条边。而这是k个节点之间实际存在的边数是E。那么这k个节点之间实际存在的边数是E与总的可能的边数$k(k-1)/2$之比就定义为节点i的聚类系数C_i,而对网络中所有节点的聚类系数取平均值,就是整个网络的聚类系数C。

3　复杂网络仿真

复杂网络可以按照不同的方式来分类,如根据节点度分布的不同,可以将复杂网络分为指数网络和无尺度网络两大类。指数网络的度分布 $P(k)$ 随度 k 呈指数衰减,ER 随机图和 WS 模型属于指数网络;无尺度网络的节点度 k 服从幂律分布,即 $P(k) \propto k^{-\gamma}$,模型属于此类网络。从生成方式上又可将复杂网络分成随机网络和确定性网络。

采用 Barabasi 和 Albert 提出的 Scale-free 模型[8]进行仿真。该模型描述了 Scale-free 网络形成的两个必不可少的机制——增长和择优链接。Scale-free 模型可以概括为:

(1)增长性。网络最初有 m_0 个节点,每一步向网络中添加一个新节点,新节点通过 $m(\leqslant m_0)$ 条新加入的连接边与网络中原有 m 个不同节点相连接。

(2)优先连接性。新节点优先与网络原有节点中出度数大的节点相连,新节点与节点 i 相连的概率 $\prod(K_i)$ 正比于该节点的度数 K_i ,为 $\prod(K_i) = K_i / \sum_j K_i$ 。其中, $\sum_j K_i$ 为网络中所有节点连接数的总和。BA 无标度网络模型根据上述公式来计算每一个节点的优先连接概率,由此得到幂律形式的网络度分布。

根据以上规则使用 MATLAB 进行仿真,网络的连接用邻接矩阵 $a_{n\times n}$ 来表示,其中 n 表示网络的节点数。当节点 i 和节点 j 有边连接时, $a_{i,j} = 1$;当节点 i 和节点 j 无边相连时, $a_{i,j} = 0$ 。

当对一个网络进行仿真分析时,它的节点数越多,得到的网络以及统计规律就越具有普遍性。由于需要存储每一对网络节点之间边的连接情况,用来存储一个网络需要的内存量可能会很大。所以采用稀疏矩阵的方法。因为小世界网络之间的连接其实是很稀疏的,表现在邻接矩阵上就是只有小部分位置是 1,大部分位置的值为 0。使用稀疏矩阵可以节省大量的内存空间,也使得对大规模网络(如万维网)的仿真更具可行性。

图 2 所示是 BA 无标度网络模型的 MATLAB 仿真结果(平均度为 6.155 6,聚类系数为 0.247 58,平均路径长度为 2.503 6)。图 3 所示是网络中节点度的分布,图 4 表示了节点度的分布概率。

4　WWW 模型的构建

将万维网中的网页作为拓扑图的一个节点,网页之间的超链接作为拓扑图的有向边。考虑到内存空间的分配情况以及搜索算法的运行时间,创建新节点时由程序自动分配全局唯一 ID 保证其可标识性,网页的 URL 作为该节点的属性存储在节点中。程序应建立一个全局 ID 分配表,随着节点的增加动态更新。节点的数据结构定义为四部分:全局 ID、网页 URL、返回指针以及 outlist 表指针。其中,outlist 表记录该网页所包含的超链接 URL。

采用邻接矩阵定义网页间的关系。如果节点 i 与节点 j 之间有边相连,对应的邻接矩阵中第 i 行第 j 列的元素为 1,否则为 0。

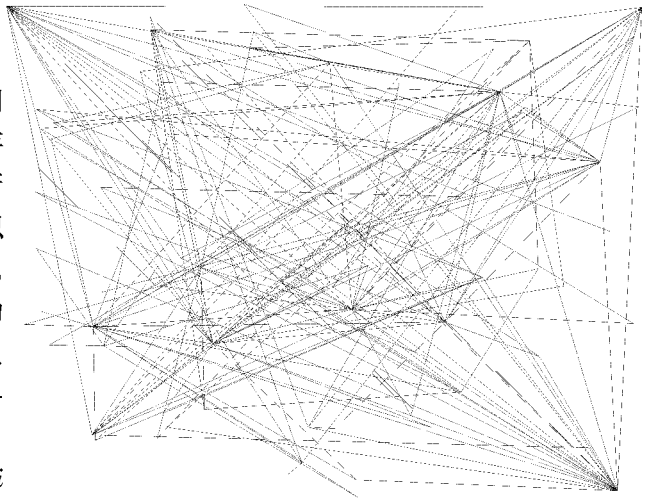

图 2　$m_0 = 10, m = 3$ 时的 BA 无标度网络图

定义 1　设 V 是一个网页集合,可以依此生成一个拓扑图 $G = (V,E)$,其中, V 表示网页节点集,若网页 $p,q \in V$,在网页 p 中存在一条超链接指向网页 q ,则在图 G 中创建一条有向边 $e = (p,q) \in E$ 。

定义 2　给定图 $G = (V,E)$,其中 $V = \{p_1, p_2, \cdots, p_n\}$,矩阵 $\boldsymbol{W}$ 是图 G 的邻接矩阵,其中,如果 (p_i, p_j) 是 G 中的有向边,则 $W_{ij} = 1$,否则 $W_{ij} = 0$ 。

由以上定义可知,万维网属于一个边权值均为 1 的有向图,要计算这个有向图的全局最短路径,经典的算法有 Dijkstra 算法和 Floyd 算法。本文拟采用 Dijkstra 算法来计算万维网的平均最短路径。

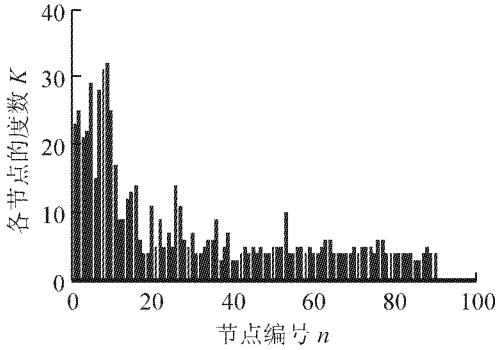

图3 各节点度的分布情况

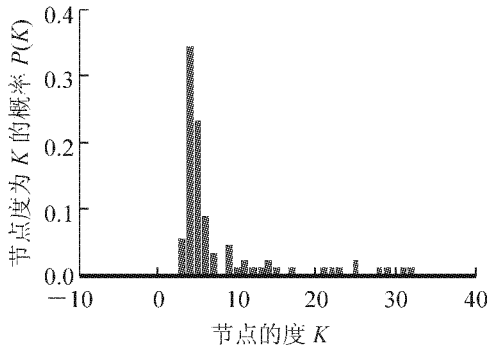

图4 节点度的概率分布情况

1959年Dijkstra提出了Dijkstra算法[9],又称为单源最短路径,所谓单源是在一个有向图中,从一个节点出发,求该节点至所有可到达节点的最短路径问题。算法的基本思想是:对网络中的每一节点赋予临时标号(源点除外),在迭代过程中不断更新这些标号。每一步,节点的临时标号表示从源点到该点的最短路径长度的上界。开始,所有节点的标号都是临时标号,每次迭代,从所有的临时标号中选取最小的,把它变成永久标号,然后重新计算拥有临时标号的各节点的临时标号,得到永久标号的节点就找到了源点到该节点的最短路径。

Dijkstra算法是图论中求解网络上两节点间最短路径的经典算法,也被认为是有效算法之一。但该算法在求解最短路径时,对两点间最短路径以外的大量节点做了计算,从而大大影响了计算速度,当网络规模较大时,尤为明显。从数学理论看,Dijkstra算法是最有效的算法之一,但在具体实施过程中,可通过各种技巧来提高实际运算效率,如通过数据结构、程序编码等,减少对点的搜索顺序和搜索量,以提高其实际计算效率。

北京大学网络实验室通过对万维网的采样得到一个包含37 482 913个网页数据集——CWT200g。若以邻接矩阵的形式存储数据,Dijkstra算法的空间复杂度为 $O(n^2)$,假设采用位矩阵的形式存储邻接矩阵,1个字节为8位,那么其用于存储邻接矩阵所需的空间为 $374\ 829\ 132^2/(8\times1\ 024^4)\approx160$ T,如此庞大的空间需求显然目前一般的PC机无法满足;其时间复杂度为 $O(n^3)$,假设计算机每秒可解决10 000个节点的最短路径计算,其所需的计算时间为 $374\ 829\ 133^3/(10\ 000^3\times60\times60\times24\times365)\approx1\ 670$ a,根本就无法实现。基于以上的分析,得出传统的Dijkstra算法难以处理这样的大规模的网络最短路径计算问题,要解决这样的大规模问题,必须设计合理的数据结构降低存储空间的需求,同时时间上要设计合理的算法尽量减少计算量,并且可考虑采用并行计算的思想。

5 路径搜索策略(PSS)

万维网上的网页数以亿计,更新速度也相当快,在构建万维网模型时务必要考虑到其动态更新的特性。旨在寻找任意两个网页之间的链接路径,并不需要将万维网所有网页都在模型中体现出来。因此,为了加快程序运行速度,节省存储空间,仅需根据输入的初始网页及其包含的超链接动态生成模型,以生成搜索树的方式获取初始网页与目标网页之间的链接路径。

(1)路径搜索算法。输入:初始网页 P_c,目标网页 P_t;输出:链接路径上的ID及其对应的URL;OPEN:算法已搜索但尚未扩展的节点集合;CLOSED:算法已扩展的节点集合;初始化:全局ID分配表填写两个表项:初始网页和目标网页。

(2)算法步骤。①把初始节点 P_c 放入OPEN表中。②若OPEN表为空,则搜索失败,退出。③读取OPEN表中第一个节点 P,将其移入CLOSED表中。④若节点 P 的ID与目标节点 P_t 的ID相同,则搜索成功,转到步骤⑦。⑤若节点 P 的outlist表为空,删除该节点,则转到步骤②。⑥若节点 P 的outlist表不为空,生成一组新节点,对这组节点作如下处理:为新节点分配ID时,首先搜索全局ID分配表中是否存在与其URL相同的表项,若不存在,分配其新的ID;对已存在于OPEN表中的节点也要删除掉,在删除之前

要比较返回初始节点的新路径与原路径,如果新路径“短”,则修改这些节点在 OPEN 表中的返回指针,使其沿新路返回;对已存在于 CLOSED 表中的节点,也作同样处理,修改其返回指针;令其余节点返回指针指向 P 后移入 OPEN 表中,转到步骤②。⑦根据节点 P 的返回指针依次逆序输出路径上各节点的 ID 及其对应的 URL,程序结束,退出。

(3)说明。①初始节点的父节点编号为空;②返回指针是父节点在 CLOSED 表中的编号,设置返回指针的目的是为了记录搜索路径,步骤⑥修改返回指针的原因是,因为这些节点又被第二次生成;③算法中对路径的长短是按路径上的节点数目来衡量的。

要计算网页间的平均路径长度,需要具有通过超链接相连接的网页节点对的个数和节点对的路径长度。对于万维网,可以将节点分为两种:一种是网站内部的节点对,一种是网站间的节点对。第一种节点的数量级一般在 10^4 以内,链接的数量级在 10^5 以内,可以在有限时间内得到路径长度的精确解。本文主要讨论第二种节点路径长度的计算。

通过对 CWT200g 数据集所有收录网站主页的链接提取处理,得到一个包含 14 858 个网站,39 393 条超链接的实验数据集 linknet。为了验证上述路径搜索策略的正确性,我们从实验集 linknet 中随机提取含有 50 个节点,93 条边的子集进行验证。输入初始节点和目标节点的编号,程序返回两节点间的路径及路径长度。比如,输入初始节点编号为 12,目标节点编号为 25,程序返回:>>起始点(12)到终止点(25)点的路径为:25 < -20 < -19 < -45 < -13 < -42 < -12;>>路径长度:6.000 00。

6　算法分析

该算法凭借网页上提供的超链接动态生成新节点并逐步扩展,搜索策略主要使用广度优先搜索,遵循从初始结点开始一层层扩展直到找到目标结点的搜索规则,而万维网上的链接指向具有一定的随机性,可能导致节点数目较多,而目标节点又处在较深层,较大的搜索量影响到算法的运行效率。可以采用初始节点和目标节点同时扩展进行搜索的方法,初始节点读取 outlist 表进行扩展操作,目标节点需读取相应的读取 inlist 表进行反向扩展,当两个方向的搜索生成同一结点时终止搜索过程。因此模型中节点的数据结构修改为下述五个部分:inlist 表指针、全局 ID、网页 URL、返回指针以及 outlist 表指针。其中,inlist 表存储所有链接到该网页的 URL。另外,OPEN 表和 CLOSED 表也应定义为二维表,分别存储两个方向上的生成结点和已扩展结点,OPEN 仍然是具有“先进先出”的队列结构。

上述程序返回结果验证了路径搜索策略的正确性,说明该策略是可行的。但对于大规模网络进行路径搜索,比如实验集 linknet,经过测试其搜索效率还是比较低的。如何提高搜索效率,是下一步重点研究的问题。

7　网络统计量的计算

按照域名的划分,将网页数据集 CWT200g 分为 13 个大类,每类数据均以其所属域名命名,再加上前面所提到的实验数据集 linknet,共计 14 套实验数据。使用 Pajek 软件[10]进行相关统计量的计算,部分统计数据如表 1 所示。

表 1　部分域名分类的统计量信息

域名	网站数	平均度 $< k >$	平均路径长度 $< d >$	网络直径 D
ac. cn	425	4.4376471	3.370 44	7
biz	97	1.958 762 9	1.010 42	2
edu. cn	1 851	5.885 467 3	4.136 54	11
gov. cn	1 652	5.244 552 1	6.032 03	20
info	91	1.956 044 0	1.264 46	2
name	79	1.974 683 5	1.000 00	1
net. cn	1 426	2.833 096	1.403 70	5
org	1 773	3.539 763 1	1.237 39	4
org. cn	927	3.007 551 2	1.977 39	4
linknet	14 858	5.302 463 3	3.996 90	13

8 结束语

小世界效应的提出对分析复杂网络具有深远的意义,米尔格兰姆的送信实验也带来两点启示:一是任何两个人都是存在联系的;二是人们可以凭借自己的交际圈寻找到这种联系。万维网也是如此,可以不去考虑网页的内容(主题等),仅依靠网页上提供的超链接去寻找任意两个网页之间的链接路径。本文提出的PPS路径搜索算法可以初步实现这一目标。当前万维网中存在的大量广告链接以及网页间相互指向的现象极大地降低了算法的运行效率。如何避免这种无效冗余的循环搜索,改进算法提高运行效率,仍是值得探讨的问题。本文提出的研究结果为建立一个通用的最佳小世界网络数据模型提供了可靠的科学依据,从而为工程信息资源的安全保存和高效检索、为提高工程信息管理中的兼容性和时效性,创建高效的元数据模型及其设计方法奠定了技术基础。

参 考 文 献

[1]S Milgram. The small world problem[J]. Psychol Today, 1967,1(1):60-67.

[2]Jeffrey Travers,Stanley Milgram. An Experimental Study of the Small World Problem[J]. Sociometry,1969,32(4):425-443.

[3]R Albert,H Jeong,A L Barabasi. Diameter of the world-wide web[J]. Nature, 1999, 401: 130-131.

[4]中国互联网信息中心. 2005年中国互联网络信息资源数量调查报告[R/OL]. [2009-10-20]. http://www.cnnic.net.cn/download/2005/20050301.pdf.

[5]D J Watts,S H Strogatz. Collective dynamics of small-world networks[J]. Nature, 1998, 393: 440-442.

[6]M E J Newman,D J Watts. Renormalization group analysis of the small-world network model[J]. Physics Letters A,1999, 263:341-346.

[7]S Boccaletti, V Latora, Y Moreno,et al. Complex networks: Structure and dynamics[J]. Physics Reports,2006,424:175-308.

[8]A L Barabasi, R Albert, H Jeong. Scale-free characteristics of random networks: the topology of the world-wide web[J]. Physica A,2000,281:69-77.

[9]EWDijkstra. A note on two problems in connexion with graphs[J]. Numerische Mathematik,1959,1: 269-271.

[10]V Batagelj. Networks/Pajek[EB/OL]. [2010-3-4]. http://vlado.fmf.uni-lj.si/pub/networks/pajek/default.htm.

Research of Small World Effect in World Wide Web

Zhao Zhengxu, Guo Yang, Liu Jiajia, Long Rui

(School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043,China)

Abstract: For long term retention, preservation and utilization of engineering information, there have been problems in compatibility between data and its host systems and problems in lead-time and efficiency for regular data check, migration and transformation, for which there has so far been no practical solutions and available satisfactory methods. This article presents the research into the small-world effect and clustering of the world-wide web and the analysis of its linkage among the web pages, exploring the path searching algorithm between the pages. It calculates the Character Path Length by the MPI-based parallel algorithm and shows that the world-wide web has the small-world effect. The research aims at a reliable benchmarking for managing large-scale and diverse engineering information and a generic small world network data model which can best cater for long-term retention and preservation and effective use of engineering data resources, therefore to enhance the data compatibility and efficiency of engineering information management via establishing a highly effective Meta data models and the related design methods.

Key words: six degrees separation; small world phenomenon; path search; parallel algorithms