

维基百科链接网络实证分析*

王辉¹, 徐晓旻², 施 佺³

(1. 上海电力学院 经济与管理学院, 上海 200090; 2. 复旦大学 计算机科学技术学院, 上海 200433; 3. 南通大学 计算机科学与技术学院, 江苏 南通 226019)

摘要: 理解维基百科词条链接网络的结构特征是深入而有效地应用维基百科的前提。基于2010年1月的数据,从度分布、权分布、宏观结构特征等角度对维基百科词条链接网络的结构特征展开实证分析。相关结果与2006年之前的维基百科词条链接网络展开对比,发现当前的维基百科网络仍然具有无标度网络特性;其宏观结构总体上满足bow-tie模型,但模型中的各组成部分比例发生了显著变化。因此,该研究对深入解析维基百科词条链接网络的结构特性具有理论和现实意义。

关键词: 维基百科; 无标度网络; bow-tie模型

中图分类号: TP391

文献标志码: A

文章编号: 1001-3695(2012)05-1637-04

doi:10.3969/j.issn.1001-3695.2012.05.009

Structural analysis on Wikipedia linkage network

WANG Hui¹, XU Xiao-min², SHI Quan³

(1. School of Economics & Management, Shanghai University of Electrical Power, Shanghai 200090, China; 2. School of Computer Science, Fudan University, Shanghai 200433, China; 3. School of Computer Science, Nantong University, Nantong Jiangsu 226019, China)

Abstract: In order to make better use of Wikipedia, it is necessary to have a clear understanding of the structural features of word entry hyperlink network of Wikipedia. This paper conducted brief experimental analysis in aspects of degree distribution, weight distribution and marco structural analysis on Wikipedia crawled on January 2010. By comparing these characteristic with previous works done before 2006, it found that the Wikipedia hyperlink network still followed power-law distribution on degree and the marco structure still obeyed bow-tie model, however, the proportion of different components defined in bow-tie model varied significantly.

Key words: Wikipedia; scale-free network; bow-tie structure

维基百科全书(www.wikipedia.org)是一个基于Wiki技术的全球性多语言百科全书网站。自从2001年1月13日上线以来至2011年5月,维基百科的英文版已有超过350万条条目,全球所有276种语言的独立运作版本共超过1700万条条目。目前,Wiki已经成为人们学习相关知识的重要途径之一。由于维基百科为开放式网络百科全书,任何个人可以对页面内容进行编辑、增加、删改、合并或分割页面,其内容涵盖面非常广泛,可以用计算机算法代替人工手段进行信息的摘要、分类、聚簇^[1]等一些列操作。这使得维基百科在词义及内容上常被作为参照物,用于为万维网等其他网页添加内容标签、消除不确定性^[2,3]。通过对维基百科的网络结构进行研究,了解其随着时间的推移而变化的过程,可以平行地将结论推广到万维网等瞬息万变且无法很好地进行记录的网络中去,从而有助于了解网络的宏观结构和发展情况;同时也可以较好地理解目前信息爆炸时代知识发展的方向和形式。

本文针对维基百科词条链接网络的结构特征展开实证分析研究。将从微观(点与边)和宏观(度分布以及总体轮廓)两个角度展开分析,与先前的相关研究结果展开比较性分析,探

讨差异形成的原因。

1 相关工作

在网络结构方面,Barabasi等人^[4]证实了真实网络度分布满足幂律分布,这类网络通常被称之为无标度网络。各类真实网络,如自治级别的互联网、万维网、电话线路、语言网络、电影演员合作关系网络、科学家合作关系以及代谢系统等都满足无标度特性,并且其负幂指数值大部分落在2~3之间。Broder等人^[5]研究了维基百科在其网络结构上与万维网相似并符合无标度网络特征及bow-tie模型。同时由于网页上具有时间戳,维基百科也比万维网在时间序列上具有网络结构的变迁特性,进而证实对于节点数较大(超过2亿个点)的万维网仍属于无标度网络,并且出度及入度均满足幂律分布。其进一步揭示了万维网的宏观结构满足bow-tie模型:大部分点属于同一个弱连通分量;而在这个弱连通分量中又存在四个主要的组成部分,如图1^[5]所示。

图1中心为一最大的强连通分量,在中心两端分别有点集只能进入这个中心或者只能从中心到达。以上三个部分形成

收稿日期: 2011-09-20; **修回日期:** 2011-10-24 **基金项目:** 国家自然科学基金资助项目(61171132);江苏省自然科学基金资助项目(BK2010280);上海市教委科研创新项目(10YS163);上海市“优青”项目(sdl09015);南通市应用研究计划资助项目(K2011003)

作者简介: 王辉(1981-),女,讲师,博士,主要研究方向为复杂系统分析(wanghuilei@gmail.com);徐晓旻(1988-),男,硕士研究生,主要研究方向为图数据库;施佺(1973-),男,副教授,博士,主要研究方向为图数据库、社会网络分析。

一个比较平衡的蝴蝶结型。在蝴蝶结的外围,还有部分点只与蝴蝶结的两端相连,只与一端相连的称为 tendril,而连接两端的则称为 tube。在这个模型中,有着很明显的平衡性:中心、蝴蝶结的两端 in 和 out、tendrils 分别约占图形的 1/4,其中心部分较大一些。

早在 2006 年 Capocci 等人^[6]证实在 2004 年时的维基百科网络符合一般万维网的两个特征:点的出度及入度的分布满足幂律分布;宏观结构满足 bow-tie 模型。同年,Buriol 等人^[7]进一步给出了 2002 年 1 月至 2006 年 1 月期间,维基百科中条目、超链接及宏观结构等的变化趋势。除了条目增加和边数增长外,还指出了在四年的过程中,各点在 bow-tie 模型的四个区域间迁移的变化趋势,如图 2^[7]所示。

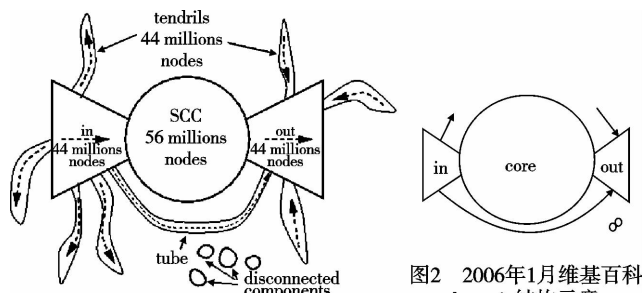


图1 Bow-tie模型

此外在理解和利用维基百科的词条链接网络的结构特征方面如社团结构分析^[8]、词条编辑合作关系结构分析^[9]也是当前研究的热点之一。

2 数据预处理

本文采用 2010 年 1 月的维基百科英文版内容,实验前对内容进行预处理,得到实验中使用的维基百科词条链接网络。下面将维基词条链接网络建模成一个有权图 $G(V, E, W)$ 。其中: V 为词条集合, $E \in V \times V$ 为词条之间的链接。需要注意的是,词条的链接是有向的,因此这里的每条边都是有向边。 $W: E \rightarrow N$ 为边上的权重函数,定义为词条之间的链接次数。

1) 数据来源 与万维网等网络的研究需要在同一个域名下通过爬取才能得到数据集不同,维基百科提供了下载服务,可以从服务器下载全部页面内容的压缩文件。与其他网络数据相比,维基百科的资料较容易得到,且内容全面没有遗漏。

2) 数据预处理 对于下载后的全部网页内容,进行以下预处理操作:

a) 将每个页面视为图的一个顶点,按数字顺序进行编号,同时记录每个页面的标题作为该页面的主题,形成如 $(A, \text{content } A)$ 的记录。由于存在重定向(同一条目旧的页面在请求时被重定向到新的页面),网络中取得的点数应该超过实际内容的条目。

b) 对于每一个网页 A ,遍历其超链接作为有向边。在网页 A 中可能不止一次链接到同一个页面 B ,以链接次数作为边的权值 W ,即生成形如 (A, X, W_{AX}) 边的记录。在页面上当一个条目的解释内容较多并且有很多方面时会产生一个目录,目录中内容全部指向同一个页面,因此会产生一种特殊的边 (A, A, W_{AA}) ,即自环。由于自环对于结构分析而言没有意义,自环将被忽略。

3 实证分析

3.1 无标度网络验证

验证 2010 年初的维基网络是否仍为无标度网络,并且就权值分布和引用情况进行实证分析。对于某个给定的图 G ,如果其度分布满足

$$P(i) = k \times \frac{1}{i^x} \quad (1)$$

其中: $P(i)$ 为任意顶点度数为 i 的概率。对于式(1),两边同时取对数得

$$\log(P(i)) = k - x \log(i) \quad (2)$$

因此验证幂律分布的基本方法为:首先统计每个出/入度数顶点的出现频率,以此作为 $P(i)$;然后对 i 和 $P(i)$ 两值同时取对数并对其进行相关性分析,通过线性拟合求出的斜率即为 x 值。

图 3 针对出度和入度分别给出了其双 log 曲线。从图 3 可以看出,对于出度和入度明显满足幂律分布,2010 年度的维基百科所形成的图仍是无标度网络。但两者的指数略微不同,对于入度其指数 x 值约为 2.31,而出度分布的指数 x 值约为 2.67,比入度稍大。对比 2004 年的维基网络,出、入度分布形态基本不变,仍然满足幂律分布,但是出度、入度分布的幂指数 x 值均增加(2004 年出入度分布指数介于 2.0 ~ 2.2^[6])。可见,在维基网络的发展过程中,图 3 中度数较小的点所占比重日益加大。

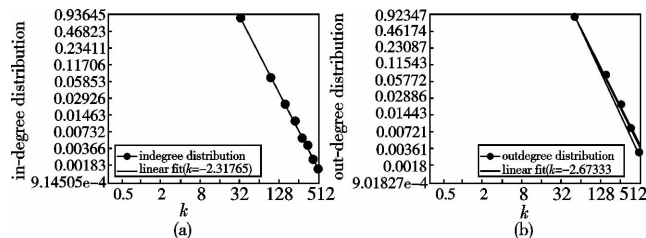


图3 2010年1月出度和入度双log曲线

对比万维网发现,当前维基网络与万维网较为相近(入度分布指数 x 值约为 2.1,出度分布指数 x 值约为 2.45^[10])。由此可见,虽然维基百科只是万维网中相对独立的一部分,并且与万维网有着较多内容上的差距,但从度分布角度而言,维基百科网络与万维网的总体情况是相似的。

在维基网络构造过程中,记录了从点 A 到点 B 的引用次数,并以此作为权重。因此除了度分布外,可以进一步研究权重分布,其结果如图 4 所示。由图 4 可知权重一般较小,在同一个页面中至多提到(引用)另一个术语 10 次。其分布同样满足幂律分布,其指数为 3.19,超出了常见无标度网络的数值(在常见的网络中,度分布的 x 值在 2 ~ 3 之间^[10])。这说明权重较小的链接占据绝大多数。

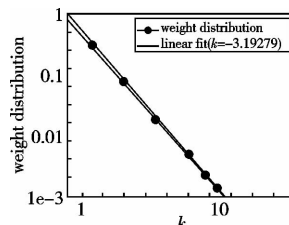


图4 权重分布

3.2 Bow-tie 结构验证

首先将维基百科形成的图视为无向图并计算连通分量,然

后在最大的分量中,将图视为有向图求出最大强连通分量,以此为基础将点进行分类,并绘制目前维基百科形成图的宏观结构。

对于弱连通分量的计算,将维基百科视为无向图,采用广度优先搜索,从一个点开始,对于每个点一旦被搜索到则加入连通分量的集合中。当以这个点开始搜索结束时,选择其他不在连通分量中的点进行广度优先搜索,直至求出全部集合。计算结果发现,已经不存在孤立的点,100%的点在同一个弱连通分量中,即在目前的维基百科中,不存在孤立的条目,任何一个条目都至少与另一个条目内容相关。

对于强连通分量(SCC),采用 Tarjan 算法^[11]。该方法基于对图的深度优先搜索,每个强连通分量为搜索树中的一棵子树;搜索时把当前搜索树中未处理的节点加入一个堆栈,回溯时可以判断栈顶到栈中的节点是否为一个强连通分量。

通过计算其中的强连通分量和出度、入度对图中的内容进行分隔,结果如表1所示。

表1 顶点分布

类别	点数	占比/%
最大 SCC	2 121 195	42.357 2
其余 SCCs	156 223	3.119 5
入度为0	2 475 325	49.428 7
出度为0	255 130	5.094 6
其他	0	0

由表1得出最大强连通分量含有2 121 195个点,约占总点数的42%,其中有98 936 239条边,约占总边数的95.3%;平均每个点有出/入边46.64条,与总体的平均边数21相比超过一倍。另一方面,这42%的点组成的强连通图包含了绝大部分的边。这说明这些点内部的引用关系是非常紧密的,而在这个强连通图之外的点,引用关系较少。

除了最大的强连通图,还有一些小的强连通图集合,含156 223个点,约占3%,与最大的强连通图相比,几乎可以忽略不计。此外,还有部分出度或入度为0的点,这些点的特殊之处在于:其只被引用或者引用了其他点,任何点无法通过它们到达其他点。这些点处于网络的两端:入度为0的点所代表的页面只能被搜索到,不能经由其他页面通过超链接进行浏览;出度为0的点代表了浏览的终点,如果不进行新的搜索或者退回,不会进入其他页面。

入度为0的点共2 475 325个,共有2 562 927条出边,平均出度约为1.035。由于这些点在同一个弱连通分量中,每个点至少有一条出边。因此,最多只有3.5%的边有两条以上的出边,基本上这些出度为0的点都只有一条出边,如词条!Atame!等,大部分为比较边缘的内容或者新事物。而出度为0的点共255 130个,经394 567条边从中间层引出,平均入度约为1.55。这部分点所代表的内容并没有引用任何其他条目,可以看出,这部分点是非常少而且也很少被其他内容提及。这部分比词条多一些边缘的内容,如!Alarma!_(album)表示唱片信息等,不再有链接出的词条。但注意到,检查2011年的维基百科时,发现其中有一些已经被加到其他词条的出边。

目前,将近一半的点入度为0,而出度为0的点在1年内出度增加,这可以归结于维基信息量的爆炸式增长。虽然从前面的初步统计可知,目前条目的增长速度已经远不如2002—2006年,但是由于基数较大,每天仍有很多新的记录或者知识被添加进入到维基百科中,而人们对这部分内容的解释最初可

能会比较简略,甚至没有制作超链接连向其他页面。由于是新生或者边缘的事物,所以不容易被其他页面的条目提及。而随着时间的推移,这些内容有一部分得到了完善,链接的词条会显著增加。

此外由表1可见,在维基百科形成的图结构中不存在此类点:出度、入度同时不为0且不在某个强连通分量中的点。由于没有此类点,在强连通分量之间或者点到强连通分量之间若有路径可达,只有两种情况:直接可达或者通过某一个强连通分量可达,当不是直接可达时,必然没有唯一路径。

1) 宏观结构模型

对于其余SCC中的点,计算其与最大强连通分量的关系,发现这部分点全部可以经由最大的一个强连通分量进入。由上文强连通分量间的链接关系和对于点的分类可以发现,不存在任何可以作为通道的点,即bow-tie模型中tube部分的顶点。

由此得到如图5所示的维基百科网络的形如bow-tie的模型。由图5可以比较明确地识别bow-tie模型中的两个部分:a)最大SCC,即中间层的最大强连通分量部分,约占42%的点;b)其余SCCs,这些连通分量构成了bow-tie模型out的一部分。

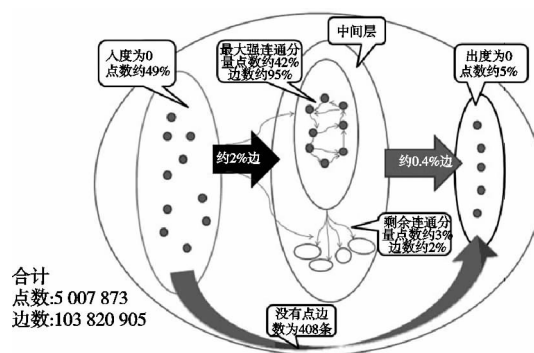


图5 维基百科构成的图的宏观结构

对于in、out的部分和tendrils部分是无法明确区分的。因为在文献[5]中采取广度优先的方式对点进行计算,如果一个点先到达的是最大强连通分量,则判定其为in的部分,若一个点先到达的是out部分,则判定为tendrils部分;类似地,从最大强连通分量到达一个点的距离较短,则认为其为out部分,若是从in部分到达此点的距离较小,则认为是tendrils部分。本文采用的是基于深度优先搜索的tarjan算法,所以无法计算此部分节点。

图5中由于不存在几个单点组成的通道,所有入度为0的点均直接到达强连通分量。因此当可达时入度为0的点到达最大强连通分量,其余强连通分量及出度为0的点的距离均为1。在此种情况下,每个入度为0的点可能属于in部分或tendrils部分。然而平均每个入度为0的点出度为1.035,因此大部分节点的平均出度为1。由于SCC比例(42%)远大于剩余连通分量所占节点比例(3%),因此可认为in部分的规模与入度为0的节点数量相当。类似地对于出度为0的点而言,仅有408条边使入度为0的边与其直接相连,因此out部分的规模与出度为0的节点数量相当。

综上近似地认为tendrils部分是可忽略的,将入度为0的点归入in的部分,而将出度为0的点与其余的强连通分量视为out部分。因此可以认为,维基百科形成的网络中,按照bow-tie模型的分类,in、SCC、out部分各占约5:4:1。

2) 宏观结构演化分析

与文献[7]所给出的2002年1月至2006年1月的维基百科网络相对比,可知在2006年之后,维基百科中 tube 和 tendrils 部分进一步缩小,直至 tube 部分不再存在,而 tendrils 部分也可忽略;同时 out 部分继续减少至目前的8%,而 in 部分在2006—2010年的四年间比例迅速上升,到全部图的将近一半,core 部分比例则逐渐缩小,到目前的约42%。

上述演化过程可归结为 in 部分数量的迅速增长。如前所述,目前信息爆炸的情况下,很可能迅速地出现很多内容较偏僻、少见,不会被引用的条目。另一方面,由于这部分内容不能由其他页面引用到,因此浏览、编辑此部分内容的人较少。人们将大部分经历用在大的强连通分量的编辑和建立其中的链接上。在这种情况下,维基百科的网络形成了一个如图6所示的结构。作为中心的最大强连通分量内容丰富、相互之间的联系紧密;仅能进入此中心的内容较为繁杂且相对独立,每条单独进入中心区域,相互之间无联系,条目数略多于中心;而仅能从中心到达的条目则非常有限,并且联系并不紧密。

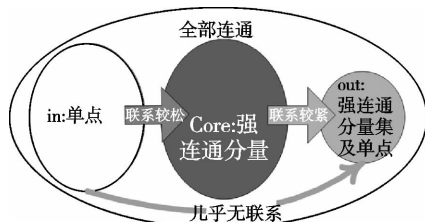


图6 维基百科2010年1月宏观结构图

与最初的 bow-tie 模型相比,可以近似地认为,到2010年1月为止,维基百科形成的图结构具有 bow-tie 模型的一些特征:全部在一个弱连通图中,存在一个大的强连通分量可以作为图结构的中心,存在只能进入或从这个中心引出的点。但是从2006年的结论已经可以看出,与 bow-tie 模型相比,作为 tube 或者 tendrils 类型的点已经可以忽略。

进一步可以得出,到2010年1月,与中间较大、两边平衡的模型相比,其宏观结构的两端出现了严重的不平衡,能够进入中心的点剧增并超过中心,两者几乎占据了全部点,而从中心引出的点较少。

4 结束语

针对维基百科词条链接网络的结构特征,验证了维基百科网络仍然满足无标度网络的特性;总体上满足 bow-tie 模型,但模型中各部分比例发生了显著变化。从上述维基百科演化趋

势可以发现,无标度网络是一个比较持久的网络特征,随着时间的推移,这一特征变化不大;而 bow-tie 特征容易受到演化的影响,2006年1月份维基百科网络所具备的比较对称的 bow-tie 形状,在很大程度上只是网络演化时的临时状态。下一步将对维基百科的编辑以及页面内容长度进行验证,并对维基百科网络微观结构的演化展开进一步研究。

参考文献:

- [1] HU Jian, FANG Lu-jun, CAO Yang, *et al.* Enhancing text clustering by leveraging Wikipedia semantics[C]//Proc of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2008: 179-186.
- [2] WU Fei, WELD D S. Autonomously semantifying Wikipedia[C]//Proc of the 16th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2007: 41-50.
- [3] KAPTEIN R, SERDYUKOV P, De VRIES A, *et al.* Entity ranking using Wikipedia as a pivot[C]//Proc of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2010: 69-78.
- [4] BARABASI A L, BONABEAU E. Scale-free networks[J]. *Scientific American*, 2003, 288(5): 50-59.
- [5] BRODER A, KUMAR R, MAGHOUL F, *et al.* Graph structure in the Web[J]. *International Journal of Computer and Telecommunications Networking*, 2000, 33(1-6): 309-320.
- [6] CAPOCCI A, SERVEDIO V D P, COLAIORI F, *et al.* Preferential attachment in the growth of social networks: the Internet encyclopedia Wikipedia[J]. *Physical Review E*, 2006, 74(3): 036116-036121.
- [7] BURIOL L S, CASTILLO C, DONA T O, *et al.* Temporal analysis of the Wikigraph[C]//Proc of IEEE/WIC/ACM International Conference on Web Intelligence. Washington DC: IEEE Computer Society, 2006: 45-51.
- [8] LIZORKIN D, MEDELYAN O, GRINEVA M. Analysis of community structure in Wikipedia[C]//Proc of the 18th International Conference on World Wide Web. New York: ACM Press, 2009: 1221-1222.
- [9] BRANDES U, KENIS P, LERNER J, *et al.* Network analysis of collaboration structure in Wikipedia[C]//Proc of the 18th International Conference on World Wide Web. New York: ACM Press, 2009: 731-740.
- [10] 车宏安, 顾基发. 无标度网络及其系统科学意义[J]. *系统工程理论与实践*, 2004, 24(4): 11-16.
- [11] TARJAN R E. Depth-first search and linear graph algorithms[J]. *SIAM Journal on Computing*, 1972, 1(2): 146-160.