第 33 卷　第 22 期　　计　算　机　工　程　　2007 年 11 月
Vol.33　No.22　　Computer Engineering　　November 2007
·网络与通信·　　文章编号：1000—3428(2007)22—0110—03　　文献标识码：A　　中图分类号：TP393

因特网知识分布研究综述

胡思康，曹元大
（北京理工大学计算机科学技术学院，北京 100081）

摘　要：因特网知识的自动获取一直是智能系统的瓶颈。而因特网网页的分布形态、网页知识的分布又在一定程度上影响知识获取及其应用技术。该文就因特网的网页分布、网页间的链接形态进行综述，指出其中存在的问题，并以此为基础，介绍了用随机行走的方法来研究因特网知识的分布。

关键词：因特网知识分布；网页超链接；随机行走

Survey of Research on Knowledge Distribution on Internet

HU Si-kang, CAO Yuan-da
(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081)

【Abstract】Automatically acquiring knowledge on Internet is always a bottleneck for intelligent systems. Web pages distribution and knowledge of Web pages distribution on Internet have influence on knowledge acquiring and its applied technology to some extent. This paper describes a survey of Web pages distribution and hyperlinks distribution between Web pages, puts forward existing problems. Method of research on knowledge distribution on Internet by random walk approach is introduced on the basis of the above description.

【Key words】knowledge distribution on Internet; hyperlinks between Web pages; random walk

目前在因特网上有数十亿个网页，其内容，特别是有意义的信息，逐渐成为以计算机为载体的一种新的知识表示方式，由此产生的因特网知识的获取、表示和应用等问题，引起了人们极大的关注。

因特网的知识，从外延来说，具有海量性、分布性、开放性、共享性、异构性、多媒体性等特点[1]。从结构上看，其分布在因特网的各个角落，网络的异构性导致知识来源不统一、知识表示不一致。从内容上看，知识的不一致性，导致对知识内容理解的歧义和不完整性。如何从大规模不规则信息中获取和组织知识，建立一套既能用于理论描述的术语，也能建立大规模计算和应用的信息表示结构，这是研究知识分布和组建的内容。由此引发的问题是：需要的知识在因特网的什么地方？它们呈怎样的分布状态？怎样才能获得这些信息？

1　因特网中的知识

1.1　因特网的发展现状

根据中国互联网络信息中心 2006 年 1 月发布的《第 17 次中国互联网络发展状况统计报告》，截止 2005 年 6 月 30 日，中国网站总数为 694 200 个。从网站数的地域分布可以看出，华北、华东、华南的网站数比例占 86.9%，东北、西南、西北网站数所占的比例不大。从网站的分布类型来看，企业类占 60.4%、个人类占 21.9%、教育科研类占 5.1%、政府类占 4.4%、商业类占 3.5%、其他类占 3.8%[2]。其中，每种类型的网站每天被访问的次数分布是[3]：企业类 2 796 次，个人类 832 次，教育科研类 3 511 次，商业类 16 240 次，政府类 2 290 次，其他类型 1 040 次。

在全国所有的 311 864 590 个网页中，静态网页占 72.7%，动态网页占 27.3%[3]。动态网页按照访问的方式分为 2 种：(1)通过点击超链接，而不用任何其他额外的输入信息；(2)必须有输入信息才能访问的网页。通常情况下，搜集的动态网页并不全面。

对网页间的超链接分析，国内各类网站的网站间链接情况分布是[3]：41.4%的网站没有链接；29.9%的网站链接了 1～5 个网站；15%的网站链接了 6～10 个网站；7.4%的网站链接了 11～20 个网站；3.6%的网站链接了 21～50 个网站；2.7%的网站链接了超过 51 个的网站(这部分数据是加权平均值)。

1.2　因特网的分布形态

文献[4]给出了整个因特网的大致形态，如图 1 所示，由以下 4 部分组成。

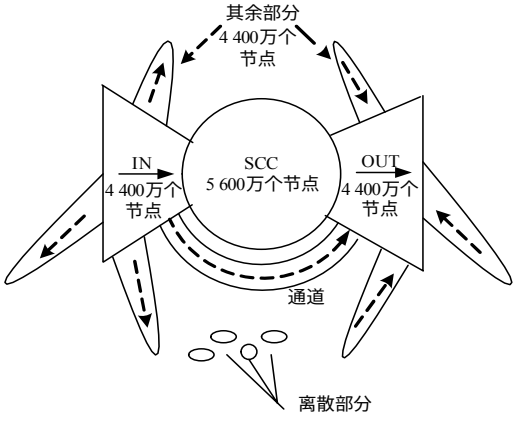

图 1　因特网形态

(1)SCC(strongly connected component)：这部分的网页彼此能互达。大多数的网站，特别是网站内部的网页，均能通

作者简介：胡思康(1975－)，男，讲师、博士研究生，主研方向：人工智能，自然语言理解；曹元大，教授、博士生导师

收稿日期：2006-12-17　**E-mail：**skhu@263.net

过链接互达。该部分占因特网的 30%。

(2)IN：这部分的网页只能从 IN → SCC，只有链出而没有链入。该部分占因特网的 24%。

(3)OUT：这部分内的网页只能从 SCC → OUT。该部分占因特网的 24%。

(4)TENDRIL：SCC 与之不能互达。该部分占因特网的 22%。

1.3 因特网中知识利用的问题

1.3.1 不同信息之间的知识融合问题

目前对搜索引擎的研究，主要关注搜索结果的查全率和查准率。但由于因特网网页变化的快速性、范围广、不确定性等因素，追求查全率很难做到。因此以 Google 中的 PageRank 算法、IBM 的 Clever 系统中的 HITS 算法、百度等搜索引擎为代表的查询算法主要关注查准率。但是目前上述算法的查准率在测试结果上仍有不足。以 Google 为例，搜索关键词“八荣八耻”，得到约 4 140 000 个相关项；以关键词“2006 世界杯”进行搜索，得到约 17 400 000 个相关项。在这么庞大的查询结果中用户能得到什么有用信息呢？再例如，今年国庆长假准备去北戴河旅游，希望能自助游且经济实惠。想知道乘坐什么交通工具，什么时间出发，路途经过多长时间等。到了目的地之后，住什么地方，有什么小吃，主要景点是什么等一系列问题。如何通过在众多搜索信息中，组合出行程路线，得到适合费用，还能给出旅游行程安排等？

为了实现上述要求，所要知道的是完成这项任务所需旅游信息在因特网上的位置，以及它们的分布状态。在这之后，才能把其中有效的分布信息以知识形式有效地存放和融合。可见，对因特网信息的利用，不仅仅是知道一个个网页、网站的独立信息，更重要的是还要知道信息网站的分布及它们之间的关系，使得搜索引擎能更好地组织利用因特网。

因特网知识的组织形式，是通过网页的组织来体现的，主要体现为以下 3 种主要组织关系：

(1)层次关系。这是网页组织的主要形式。网站内的网页或者按照层次关系，或者按照组成关系，或者按照关联关系组织。

(2)平行关系。在这种类型中，网页具有平行关系的若干子网页组成。

(3)混合关系。在这种类型中，网页之间是网状关系，网页中各元素彼此关联。

1.3.2 仅对静态网页分析的问题

无论是文献[5]中利用基于链接分析搜索算法 HITS，还是文献[6]中用二分有向图的方法对互联网上的社区给出了一种明确的划分，这些类型的研究都是在对静态资料分析的基础上，利用超链接进行分析随机行走。这些分析中还存在一些不足，主要体现在两方面：

(1)网页之间动态变化关系

多数的研究对静态网页分布的研究较多，大多完全是研究静态网页分布(网页之间的超链接)，而对网页之间链接关系的动态变化、网页静态连接和动态连接之间的关系研究较少。有理由相信，无论网页之间的关系如何发生改变，同一知识源的网页间的链接概率是很大的。也就是说，当一个网页 A 的超链接发生改变时，可能是以下几种情况之一：

1)网页 A 被删除；

2)网页 A 的链入被删除，成为孤立网页；

3)网页 A 的链出被删除，成为终止网页；

4)网页 A 的超链接被修改，但是仍链接到其他信息相关的网页/网站上。

(2)网页间链接的实际使用情况

目前较多的是研究网页之间链接关系客观存在情况，而研究这些链接实际被利用的情况却很少有人关注。针对因特网分析来说，目前的研究只涉及信息量分布研究，而没有考虑因特网的知识量分布。

2 因特网的随机模型

2.1 随机行走

由于网页中的超链接反映作者产生该链接的思想意图，带有一定的主观性，因此研究的随机行走不是严格的随机过程。另外，网页的链接统计显然是随机时间而变化的。从直观上说，随机行走是对在每一个随机方向上采取连续行走的形式化。网页之间的超链接，使得因特网构成了一个连通图。把网页(或整个网站)间的一次点击看作随机一跳，针对因特网上的随机行走，给定了以下部分：

(1)静态平面网格。用于建立随机行走之前的网络连接结构；

(2)动态平面网格。用于在用户点击页面时建立随机行走路径；

(3)随机行走规则。网页页面上定义的超链接反映了创建页面时跳转的方向；

(4)随机行走概率。通过跳转超链接所占频度来定义网页页面间随机行走概率，由式(1)定义：

$$\text{随机行走概率} \Rightarrow p(CP|OP)*\frac{\text{同一跳转链接数}}{\text{总跳转链接数}} \tag{1}$$

其中，$p(CP|OP)$ 表示从其他网页跳转到当前网页的概率。

2.2 随机行走示例

根据以上描述来分析随机行走在二维平面网格上的坐标过程[7~11]，并结合因特网特征网页链接特征，有如下假定，并采用了 3 类随机行走模型：

(1)假定网格中的节点彼此等价，整个网格链接无向、无权重、等价。

(2)起始位置和条件。起点在坐标原点(0,0)，行走区域是二维平面网格(不包括 x 的负半轴，但行走时可以跨越)，终点在位置(i,j)。

(3)相邻位置之间的一跳，距离为常量，具体值由所定义的一跳集合来定义。一跳集合 σ 分为 2 类：

1)水平随机行走：$\sigma=\{(0,n),(n,0),(0,-n),(-n,0)\}$，$n\in\mathbb{N}$，不失一般性，$n$ 取 1，如图 2(a)。

2)对角线随机行走：$\sigma=\{(n,-n),(-n,n),(n,n),(-n,-n)\}$，$n\in\mathbb{N}$，不失一般性，$n$ 取 1，如图 2(b)。

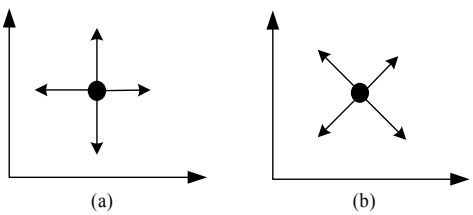

图 2 二维平面网格上的随机行走

2.3 分布和知识分布的随机规律

用随机行走模型，研究因特网用户查询流的分布和发展规律；研究网上知识被充分利用的程度以及未被充分利用的原因及其提高其利用率的有效途径；研究网页分布和知识分布的随机规律。实际上，因特网上浏览流、查询流的分布，总体上反映的是因特网知识的分布，因为浏览、查询的过程，

就是力求给出满足用户所需知识的网页链接情况，所以浏览流、查询流的分布就落在知识分布图中。以概率论为工具，构建互联网的网络动态模型，为了使该模型确能描绘因特网环境的属性，定义模型的特征如下：

(1)新创建网页/网站的链接应该以较大的概率指向那些具有较高链入数的网页。

(2)如果存在一个超链接($A\rightarrow B$)，从网页 A 出发的其他链接所指向的其他网页 C，应该包含网页 B 的有关信息，或者包含与网页 B 属于同一知识源的其他信息。

(3)如果存在一个超链接($A\rightarrow B$)，指向网页 B 的其他链接所在的网页，应该包含网页 A 的有关信息，或者包含与网页 A 属于同一知识源的其他信息。

(4)如果删除一个网页 A，则会降低从网页 A 出发链接($A\rightarrow B$)到的所有网页的被访问概率。

(5)如果删除一个网页 A，则会降低链接到网页 $A(A\rightarrow B)$ 的所有其他网页的被访问概率。

其中，被访问概率由式(2)确定：

$$P(A)=\sum_{i=1}^{m}WI_i(A)+\sum_{j=1}^{n}WO_j(A)+Q(A) \tag{2}$$

其中，$P(A)$是网页A被访问概率，它与网页A的每个链入度权值$WI_i(A)$、链入度m和链出度$WO_j(A)$、链出度数n相关，还与网页的质量Q(A)相关。这里的网页质量，指的是网页内容与所属知识源的相关度，由下面的式(5)定义。

对于网页的链接权值，定义了 3 种不同的计算方法：

(1)如果网页 A 到网页 B 有链接，且链接标记 href 周围出现越多的、与网页 A 的主题相关的内容，该链接权值就越高。与标记 href 周围文本相关的权值计算由式(3)定义：

$$W(A\rightarrow B)_{href}=\sum_{H\in HS}\frac{NUM_{href}}{NUM_A} \tag{3}$$

其中，HS 是网页 A 到网页 B 的链接集合。

对于式(3)根据不同的应用背景和结构，有相应的变形。对于建立了关键词表的系统来说，可以使用式(4)的文本向量夹角余弦来定义：

$$W_{SIM}(V_{href},V_A)=\frac{\sum_{i=1}^{n}V_{href_i}\times V_{A_i}}{\sqrt{\sum_{i=1}^{n}V_{href_i}^2\sum_{i=1}^{n}V_{A_i}^2}} \tag{4}$$

式(3)和式(4)显示了链接与包含链接的网页之间的紧密关系，而与链接到的网页 B 的主题无关。

(2)如果网页 A 到网页 B 有链接，且链接到的网页 B 质量越高，该链接权值就越高。网页质量计算由式(5)定义：

$$Q(d)=\sum_{D}\frac{NUM_d}{NUM_D} \tag{5}$$

其中，NUM_d表示网页d的关键词集；D表示同类网页的关键词集合。

(3)与网页质量权值计算由式(6)定义：

$$W_{\mathrm{QUILITY}}(A\rightarrow B)=\sum_{D}NUM_{href}\times Q(B) \tag{6}$$

式(6)显示了链接与所链接的网页 B 之间的紧密关系，且与网页 A 的链接文本主题相关。

3 结束语

网页无论是数量还是内容都在不断变换。如何从中找出隐藏的规律，如何分析和探究网页分布结构，从而使得这些网页信息能够更好地被利用，是笔者研究因特网知识分布的初衷。网络结构的分布，网页包含的信息由稀疏到密集的转变过程，在一定程度上反映了人们对信息集结、归纳的过程。对因特网结构和知识分布的研究，为使用类自然语言理解思想和技术获取网页文本知识奠定了良好的基础。

参考文献

1 陆汝钤. 研究知识科学, 发展知识工程, 推进知识产业[R]. 北京: 中国科学院计算技术研究所, 2003.

2 中国互连网络信息中心. 中国互联网络发展状况统计调查[EB/OL]. (2006-01). http://www.cnnic.net.cn/index/0E/00/11/index.htm.

3 搜狐网. 第三次中国互联网络信息资源数量调查报告[EB/OL]. (2004-08). http://www.sohu.com.

4 Kumar R, Raghavan P, Rajagopalan S, et al. The Web as a Graph[D]. Providence, RI: Computer Science Department, Brown University, 1999.

5 Gibson D, Kleinberg J, Raghavan P. Inferring Web Communities from Link Topology[C]//Proceedings of the 9th ACM Conference on Hypertext and Hypermedia. 1998.

6 Kumar S, Raghavan P, Rajagopalan S, et al. Trawling Emerging Cyber-communities Automatically[C]//Proceedings of the 8th ACM-WWW International Conference, Toronto. 1999.

7 钱敏平, 龚光鲁. 应用随机过程[M]. 北京: 北京大学出版社, 1998-10.

8 Mireille B M, Schaeffer G. Walks on the Slit Plane[M]. Berlin/Heidelberg: Springer, 2002-11: 305-344.

9 Mireille B M. Walks on the Slit Plane: Other Approaches[J]. Advances in Appled Mathematics, 2001, 27(2): 234-288.

10 Mireille B M, Schaeffer G. Walks Confined in a Quadrant Are Not Always Definite[EB/OL]. (2002-11-27). http://arXiv.math.com/0211432V1.

11 Mireille B M. Counting Walks in the Quarter Plane: Kreweras' Algebraic Model[J]. Annals of Appled Probability, 2005, 15(2): 1451-1491.

(上接第 109 页)

5 Ou Yang. Measurement Research of Dependency Relations among Classes Based on Object-oriented System[J]. 计算机科学, 2004, 31(2).

6 Fang Fei. An Approach to Object-oriented Software Regression Testing[J]. Journal of Software, 2001, 12(3).

7 萨师煊, 王 珊. 数据库系统概论[M]. 3 版. 北京: 高等教育出版社, 2000.

8 Jones C. Applied Software Measurement[M]. [S. l.]: Mcgraw-hill, 1986.

9 Fenton N E. Software Metrics: A Rigorous Approach[M]. New York: Chapman & Hall, 1991.

10 Chidamber R, Kemerer F. A Metrics Suite for Object-oriented Design[J]. IEEE Trans. on Software Engineering, 1994, 20(6).