万维网的链接结构分析及其应用综述*

王晓宇, 周傲英+

(复旦大学 计算机科学与工程系,上海 200433)
(复旦大学 智能信息处理开放实验室,上海 200433)

Linkage Analysis for the World Wide Web and Its Application: A Survey

WANG Xiao-Yu, ZHOU Ao-Ying+

(Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China)
(Intelligent Information Processing Laboratory, Fudan University, Shanghai 200433, China)
+ Corresponding author: Phn: 86-21-65643503, Fax: 86-21-65643503, E-mail: xiaoyuwang@fudan.edu.cn
http://www.cs.fudan.edu.cn
Received 2002-08-22; Accepted 2003-04-21

Wang XY, Zhou AY. Linkage analysis for the World Wide Web and its application: A survey. *Journal of Software*, 2003,14(10):1768~1780.
http://www.jos.org.cn/1000-9825/14/1768.htm

Abstract: Up to now, the World Wide Web (WWW) grows into a large hyperlinked corpus with more than 800 million pages and 5 600 million hyperlinks. Moreover, it is obviously impossible that any global 'planning' can be imposed on the creation of such a corpus. This brings some challenges to many research fields on the World Wide Web. On the other hand, the hyperlinked Web pages in the networking environment can be a very rich information source for daily or business use, provided people have effective means for understanding the Web. Linkage analysis is playing more and more significant role in many fields on the World Wide Web. Recent advances about the relevant research and application of linkage analysis of World Wide Web are presented in this paper. In particular, some results and achievements about linkage analysis and its applications on Web searching, Web community discovery and the Web modeling are surveyed here.
Key words: linkage analysis; World Wide Web; Web searching; knowledge discovery

摘 要: 当今万维网的规模已经快速发展到包含大约 80 亿个网页和 560 亿个超链接.此外,对万维网的创建进行全局规划显然是不可能的.这些都对万维网的相关研究提出了挑战.另一方面,互联网环境下通过超链连接起来的网页,为人们的日常和商务用途提供了非常丰富的信息资源,但前提是必须掌握有效的办法来理解万维网.链接结构分析在万维网的很多研究领域起着越来越重要的作用.全面介绍了万维网链接分析方面的最新研究进展和应用情况,对链接分析在 Web 信息搜索、万维网潜在社区发现及万维网建模等方面的研究进展和实际应用进行了综述.

* Supported by the National Natural Science Foundation of China under Grant No.60003016 (国家自然科学基金); the China Cross-Century Talent Raising Program of MoE under Grant No.2000-82 (教育部跨世纪优秀人才培养计划); the China Young Teacher Fund of Fok Ying Tung Education Foundation under Grant No.81062 (霍英东教育基金青年教师基金)

第一作者简介: 王晓宇(1975－),男,安徽濉溪人,博士,主要研究领域为人工智能,互联网环境下的数据搜索.

关键词: 链接分析;互联网;Web 搜索;知识发现
中图法分类号: TP393 **文献标识码**: A

随着互联网的不断发展,人们越来越多地在互联网上发布和获取信息.Web 已经成为信息制造、发布、加工和处理的主要平台.传统的互联网应用技术大多是基于文档内容的,与经典的信息检索技术和数据库技术有着密切的联系.但是,互联网中特有的许多问题,诸如超大规模的非结构化文档数量、良莠不齐的网页质量、包含在文档中的大量多媒体信息,甚至相当含糊或不规范的用户查询表示等,都使得经典的信息检索技术和数据库技术在互联网环境中很难有效地应用.

另一方面,互联网又包含了传统数据环境所没有的另一种丰富信息,即互联网的超链接拓扑结构.网页间的超链接一方面引导网页浏览的过程,另一方面也反映了网页创建者的一种判断,即有理由认为,如果网页 A 存在一条超链接指向网页 B,那么网页 A 的作者是认为网页 B 包含了有价值的信息.因此,充分利用互联网的链接结构信息对互联网应用技术的研究将具有极为重要的意义.事实上,越来越多的学者已经开始致力于这方面的研究,总体来说主要包括以下 3 个方向:

- 链接结构分析在 Web 信息搜索中的应用;
- 链接结构特征与互联网中出现的潜在社区之间的关联;
- 链接结构在理解互联网自身属性特点和成长模式方面所处的地位和作用.

本文第 1 节较为详细地介绍已有的一些基于链接分析的主题提取算法.这部分内容不仅包括了经典的 HITS 算法和 Google 中的 PageRank 算法,同时还介绍了一些重要的衍生算法,并从理论和应用的角度对这些算法进行了比较.第 2 节介绍互联网社区研究的意义以及已有的两种互联网社区发现技术.第 3 节展示了在互联网结构图分析与建模研究方面正在进行的一些初步探索.第 4 节简要介绍链接分析在其他超文本检索研究及网页智能爬取方面的一些应用现状.第 5 节探讨链接结构研究将来可能的研究方向.

1 主题提取的模型与算法

通过搜索引擎查找与某个主题相关的网页非常容易,但是,假如查询是一个相对比较广泛的主题,那么搜索引擎通常会返回成千上万的条目.尽管从某种意义上说,这些内容大多是和主题相关的,但是它们的价值程度却千差万别.而且,对网页的价值判断本身又是一个非常主观的过程,许多因素都会影响这种价值的判断,诸如站点或网页的组织和形式、信息的质量甚至独特性等等.因此,越来越多的方法开始考虑利用链接结构的信息[1].由于链接被创建的过程本身就包含了人的判断,充分而有效地利用这些信息将从很大程度上有助于这个问题的解决.

PageRank 算法和 HITS 算法是两种影响相当广泛的链接分析算法.但是,深入的研究表明,它们仍存在一些明显的缺陷[2~4],因此许多学者在此基础上又提出了一些衍生的算法[2,3,5~8],其中包括 IBM Almaden 实验室的 CLEVER 系统、Compaq 系统研究中心的 Web Archaeology 项目以及我们提出的 STED 算法.

1.1 PageRank算法

PageRank 算法[9]是最早并且最成功地将链接分析技术应用到商业搜索引擎中的算法.它的基本出发点是试图为搜索引擎所涵盖的所有网页赋予一个量化的价值度.每个网页被量化的价值通过一种递归的方式来定义,由所有链接向它的网页的价值程度所决定.显然,一个被很多高价值网页所指向的网页也应该具有很高的价值.这种规则可以用一种随机网上冲浪(surfer)的模型来描述.具体来说,如果假设冲浪者跟随链接进行了若干步的浏览后转向一个随机的起点网页又重新跟随链接浏览,那么一个网页的价值程度值就由该网页被这个随机冲浪者所访问的频率所决定.

这个过程也可以理解成一个 Markovian 过程,每个网页是一个状态,从一个网页跟随链接浏览到另一个网页可以被看作是一个状态的迁跃,所有这种迁跃的概率是相同的.但是,考虑如果存在一类网页,这类网页中不包含任何指向其他网页的链接,那么这种网页将成为沉积(sink)网页,并使得上述这种迁跃的过程在沉积网页上

永远终止.解决这个问题的方法很简单,假如一个随机冲浪者遇到了这种沉积网页,那么他可以随机地挑选另一个网页并继续他的浏览.为了对那些不是沉积的网页也一视同仁,这种类型的随机迁跃应该能以相同的概率在任何一个网页上发生.下面是整个过程的形式化表达,并由此可以为每一个网页计算其价值度 PR:

$$PR(i)=d\cdot D(i)+(1-d)\sum_{j\to i}[PR(j)/N(j)], \quad (1)$$

其中$\sum_{j\to i}$作用于所有链接向网页 i 的网页 j,$N(j)$表示链接向网页 i 的所有网页 j 的总数.冲浪者以概率 d 不再跟随链接浏览而重新从 $D(\cdot)$分布中挑选一个网页重新开始浏览,以 $1-d$ 的概率继续从当前网页中跟随一个链接浏览.

根据这种方法对网页排序以后,搜索引擎 Google 就可以决定以一种什么样的顺序将结果返回给查询用户.

1.2 HITS算法

不同于 Brin 和 Page 的 PageRank 算法,Kleinberg[10]提出了一种更为完善的衡量网页重要程度的度量.他认为网页的重要程度是与所查询的主题相关的.在 HITS 算法模型中,Kleinberg 提出了权威性网页(authority)的概念.互联网上一个广义的主题包含有大量显著的权威性网页,这些权威网页从链接结构的角度来看应该是被大量的超链接所指向的,也可以说是被大量的网页作者所认可的.然而仅通过这种计算链入数目的机制来描述互联网环境中网页的权威性在实际中仍会有很多问题.在很多情况下,同一主题下的权威网页之间并不存在相互的链接(相互间并不"认可").例如,"Microsoft"和"Netscape"虽然都是浏览器主题中的权威站点,但它们却并不存在相互的链接.然而,它们通常同时被一些不知名的网页所共同指向.Kleinberg 称这种网页为中心性网页(hub),它们指向多个主题相关的权威网页.通过这两种不同类型的网页(权威网页和中心网页),链接结构可以描述为它们之间的一种依赖关系:一个好的中心性网页应该指向很多好的权威性网页,而一个好的权威性网页则应该被很多好的中心性网页所指向.

基于以上这种链接结构描述的概念,可以定义一种区分网页价值程度的度量.具体来说,首先利用一个传统的文本搜索引擎(例如 AltaVista)获取一个与主题相关的网页根集合(root set).然后向根集合中扩充那些指向根集合中网页的网页和根集合中网页所指向的网页,这样就获得了一个更大的基础集合(base set).假设最终基础集合中包含 N 个网页,那么对于 HITS 算法来说,输入数据就是一个 $N\times N$ 的相邻矩阵 A,其中如果网页 i 存在一个链接到网页 j,则 $A_{ij}=1$,否则 $A_{ij}=0$.

HITS 算法为每个网页 i 分配两个度量值:中心度 h_i 和权威度 a_i.设向量 $a=(a_1,a_2,\ldots,a_N)$代表所有基础集合中网页的权威度,而向量 $h=(h_1,h_2,\ldots,h_N)$则代表所有的中心度.最初,将这两个向量均置为 $u=(1,1,\ldots,1)$.操作 $In(a)$使向量 $a=A^Th$,而操作 $Out(h)$使向量 $h=Aa$.反复迭代上述两个操作,每次迭代后对向量 a 和 h 范化,以保证其数值不会使计算溢出.Kleinberg 证明经过足够的迭代次数,向量 a 和 h 将分别收敛于矩阵 A^TA 和 AA^T 的主特征向量.通过以上过程可以看出,基础集合中网页的中心度和权威度从根本上是由基础集合中的链接关系所决定的,更具体地说,是由矩阵 A^TA 和 AA^T 所决定.

1.3 PageRank算法和HITS算法的进一步探讨与比较

PageRank 算法实质上是一种通过离线对整个互联网结构图进行幂迭代的方法.PageRank 所计算出的价值度的值实际上就是互联网结构图经过修改后的相邻矩阵的特征值.对这些值的计算有非常有效的方法(事实上,仅需要若干次的迭代计算即可以得到),因此能够很好地应用到整个互联网规模的实践中.这种方法的另一个主要优点是所有的处理过程都是离线进行的,因此不会为在线的查询过程付出额外的代价.但是,PageRank 算法也同样存在一个显著的问题,即价值度的计算是不是针对查询的.对于某个特定主题的查询,在返回结果中一些与主题无关的"强壮"网页将会排在较前的位置.比如,PageRank 会把网页 excite.com 网页排在 city.net/counties/greece 的前面,因为 excite.com 显然比 city.net/counties/greece 具有更大的链入数目.当查询是 Greece 时,excite.com 将会在查询结果中比 city.net/counties/greece 具有更高的价值度.当然,excite.com 可以通过文本的分析而被预先剔除掉,但是这类问题对 PageRank 算法的影响则有必要作更进一步的研究.比如,PageRank 的这种迭代过程如果只作用在特定查询的子图上,是否还会产生和作用在全局网图上同样的排列呢?

HITS算法在概念的定义上比PageRank算法多提出了一个中心性网页的概念.通过中心网页和权威网页的相互作用,HITS算法更好地描述了互联网的一种重要组织特点:权威网页之间通常是通过中心网页而彼此发生关联的.HITS算法和PageRank相似,也是通过迭代的方法计算相邻矩阵的特征向量.但HITS算法所针对的不是整个互联网结构图,而是特定查询主题的互联网子图.规模上的极大减小可以使HITS算法的迭代收敛速度比PageRank要快得多.但因为与查询相关,所以查询过程需要考虑排序的代价.

另外,除非为HITS算法中所考虑的链接赋予适当的权值,否则,相邻矩阵的主特征向量并不能反映最合理的网页价值度排列(参见第1.5节).并且,即便对子图中的边赋予了适当的权重,如果子图的相邻矩阵是一个可约减的矩阵(例如图中有多个不连通的部分),那么很多有价值的网页仍将无法在主特征向量中得到体现(参见第1.6节).更为严重的是,在对很多广义主题进行查询时,HITS算法会错误地将许多与主题无关的网页赋予很高的价值度.例如,当查询“电影奖”时,得到的结果却是许多电影公司的主页.这是因为和“电影奖”有关的网页通常会链接向电影公司的主页,由于电影公司主页的商业性,大量的链接会发生在这些公司主页之间,从而错误地诱导了HITS算法.这种现象通常被称为主题漂移(topic drift).

最后,应该注意到HITS算法所作用的查询子图是根据查询关键词在线构造的.通过常规的方法将无法满足在线查询响应时间的要求,但是,如果借助专用的连接服务器(connectivity server)[11],查询子图的构建时间将是毫秒级的.

1.4 ARC和CLEVER系统

IBM Almaden实验室开发了ARC(automatic resource compilation)系统[5]和CLEVER系统[8].虽然两个系统的目的不同(ARC用于资源的半自动编辑,而CLEVER用于互联网的搜索),但它们都是以HITS算法为核心,并试图通过增加对网页内容信息的利用来克服HITS算法的主题漂移.

改良的算法考虑在链接(href)周围的文本内容会较大程度地反映链接所指向的网页的内容.如果这个链接周围的文本出现了查询的主题,那么可以更加确信链接所指向的网页也是与查询主题相关的.剩下的问题就是如何将这种确信反映到HITS算法的迭代过程中.

基本的思想是为每个链接分配一个权值$w(p\to q)$,如果从p到q的href周围出现的与主题相关的文本越多,那么这个链接的权值也就越高.与HITS算法一样,迭代过程也始于两个向量h和a,其分量的初值均设为1.但与HITS算法不同,相邻矩阵的构造不再是布尔型矩阵.矩阵W的每一项对应于一对网页,如果它们之间存在链接,则$W_{pq}=w(p\to q)$,反之$W_{pq}=0$.迭代过程的每一步可以表示为$a=Wh;h=W^Ta$.矩阵计算的理论可以证明这个过程是收敛的.

那么,如何给出每一个链接的权值呢?必须解决两个首要的问题:第一,考虑“在href周围的文本”,但是“周围”精确的含义是什么?第二,如何将href周围的文本映射成一个量化的权值.Chakrabarti[5]的方法是将href的左右看作是一个B字节的窗口(包括〈a href=“…”〉和〈/a〉之间的文本).B是一个由实验决定的参数.设$n(t)$表示在窗口中出现的与主题相关的词(用户查询的关键词)的频度.那么,链接权重的定义可以表示为$w(p\to q)=1+n(t)$.这种定义使得矩阵W中的很多项都大于1.但是,通过每次迭代后的范化操作,这些项还是会维持在很小的数值上.

1.5 加权(weighting)和修剪过滤(outlier filtering)方案

Compaq系统研究中心的学者[3]在深入研究HITS算法后也发现,该算法在很多情况下的表现并不令人满意.他们指出导致HITS算法不足的主要原因有以下两点:

(1) HITS算法假设的前提是,不同的链接反映出不同作者的判断.但是在很多情况下,某个站点的很多网页会同时指向另一个站点的某个网页,这将导致第1个站点的那些网页的hub值的增加和第2个网站的那个网页的authority值的增加.反之,第1个网站上的某个网页指向另一个网站上的很多网页,也将会导致同样的情况.但是,这些不同的链接仍然只是代表一个组织或某个人的判断,所以,不加区分地对待它们导致了不应有的权值增加.

(2) 通常在HITS算法所构造的查询子图中会包含有与查询主题无关的网页.如果这些无关的网页周围存在很多的链接,那么,主题漂移的情况就有可能发生,即那些最权威的和最中心的网页可能和原来的查询主题没

有关系.

为了解决第(1)个问题,Bharat等人[3]提出了对链接加权的方法,不同于ARC和CLEVER系统中的权重方案.他们认为一个站点内的很多网页或者一个网页对另一个网站内的网页所产生的贡献程度应该是一样的.假设第 1 个站点内有 k 个网页指向第 2 个网站的一个网页,那么给每一条链接分配一个权威性权重(authority weight)1/k.这个权在计算网页的权威性时将会被用到.同样,假设第 1 个网站的一个网页有 l 条链接指向第 2 个网站的 l 个网页,那么给每一条链接分配一个中心性权重(hub weight)1/l.修改后的算法迭代过程如下:

$$a(i)=\sum_{(i,j)\in BaseSet}h(j)\times auth\_wt(i,j), \tag{2}$$

$$h(i)=\sum_{(i,j)\in BaseSet}a(j)\times hub\_wt(i,j). \tag{3}$$

Bharat等人将传统的信息检索技术结合到链接分析的算法中,以克服HITS算法的主题漂移问题.通过对网页的内容分析,可以预先将查询子图中的无关网页剔除出去.他们采用近似度的概念来衡量一个网页和查询主题的相关程度.由于查询主题要比查询本身更为广泛,因此,仅用查询关键词的匹配来计算相似度是不够充分的.Bharat 等人用整个根集合来定义所查询的主题.具体来说,他们将所有根集合中的网页内前 1 000 个词连接起来,作为所查询的主题 Q.网页 D_j 与查询 Q 的相似度定义由如下公式给出:

$$Similarity(Q,D_j)=\sum_{i=1}^{t}(w_{iq}\times w_{ij})/(\sum_{i=1}^{t}(w_{iq})^2\times\sum_{i=1}^{t}(w_{ij})^2)^{1/2}, \tag{4}$$

其中:

$w_{iq}=freq_{iq}\times IDF_i$;

$w_{ij}=freq_{ij}\times IDF_i$;

$freq_{iq}$=词 i 在查询 Q 中出现的频率;

$freq_{ij}$=词 i 在网页 D_j 中出现的频率;

IDF_i=词 i 在整个互联网中反比文档频数(inverse document frequency)的估计.

1.6 基于相似度分析模型(STED)的算法

在深入分析上述算法的基础上,我们提出了一种基于相似度分析模型的主题提取算法[4].借鉴文献引用分析(citation analysis)的思想[12],基于链接分析的主题提取过程可以通过如下的一种相似度模型来加以描述.具体来说,给定一个基础集合 T\{1,2,…,n\},为其中的网页从 1 到 n 加以编号.为了描述链接的结构信息,我们为 T 中的每个网页定义两个向量 v^{in} 和 v^{out},如果存在网页 i 到该网页的一个链接,那么向量 v^{in} 的第 i 个分量为 1,否则为 0.同样地,如果存在该网页到网页 i 的一个链接,那么向量 v^{out} 的第 i 个分量为 1,否则为 0.这样,两个网页间的某种相似性度量可以表示为它们向量的内积 $Similarity^{in}(i,j)=v_i^{in}\cdot v_j^{in}$, $Similarity^{out}(i,j)=v_i^{out}\cdot v_j^{out}$.容易看出,这两种相似度实际上分别描述了两个网页共同被其他网页所链接的数目以及两个网页共同链接其他网页的数目.

基于上述两种相似度的定义,可以为整个基础集合构造两种相似度矩阵 S^{in} 和 S^{out},其中 $S^{in}(i,j)=Similarity^{in}(i,j)$,而 $S^{out}(i,j)=Similarity^{out}(i,j)$.相似度矩阵的 k 次幂矩阵中的项可以被视作网页间的 k 阶相似度,即 $Similarity_k^{in}(i,j)=(S^{in})^k(i,j)$和 $Similarity_k^{out}(i,j)=(S^{out})^k(i,j)$.深入的研究分析[4]表明,HITS 算法第 k 次迭代后,网页 i 的权威值和中心值分别为 $a(i)=\sum_{j(j\neq i)}Similarity_k^{in}(i,j)$;$h(i)=\sum_{j(j\neq i)}Similarity_k^{out}(i,j)$.

如果从图论的观点出发,可以发现(i,j)的 k 阶相似度实际上是 k 个一阶相似度的乘积.这 k 个一阶相似关系构成了从 i 到 j 的一个传递路径.理论分析和实验表明,正是这种两两之间的传递经常扭曲了 i 和 j 之间的相似关系.为了解决传统的相似性度量只能描述两两之间关系的问题,我们借助于关联规则挖掘技术,提出了一种更为一般的相似度定义.对应于互联网的链接结构,可以将基础集合中的所有网页看作一个条目集合(itemset),一个网页的所有链入链接或链出链接可以被看成是一个事务.通过关联规则算法发现的频繁项集(frequent item set)则对应于一组具有共同链接(co-citing or co-cited)网页的网页集合.值得注意的是,通过频繁项集所捕捉的是两个或两个以上网页间的关系.这种关系的度量ε由下式给出:

$$\varepsilon(I)=(\sum_{t=1}^{k}\mu_t)/k. \tag{5}$$

其中,μ_t 为频繁项集 I 中第 t 个关联规则的置信度(confidence),k 为频繁项集 I 中所包含的基本关联规则(essential rules)数.

为了仍然能够利用简单的迭代操作,需要找到一种方法,将上述这种多个网页间关系的描述映射到传统的两两相似度的概念上.具体的映射由下式给出:

$$\varsigma(i,j)=\sum_{\{I|i,j\in I\}}\varepsilon(I). \tag{6}$$

根据新的相似度 $\varsigma(i,j)$,可以构造新的相似度矩阵 S.和 HITS 算法相似,通过若干次的迭代过程,我们可以得到基础集合中所有网页的中心值和权威值.迭代过程的第 k 步可以表示为

$$a_k=S^{\text{in}}a_{k-1}, \tag{7}$$

$$h_k=S^{\text{out}}h_{k-1}. \tag{8}$$

在第 1.3 节我们曾提到,当查询子图的相邻矩阵是可约减的矩阵时(查询子图包含了若干个不连通的子图),通过类似HITS算法的迭代过程所得到的结果会丢失一些很好的权威性网页和中心性网页.实验证明,当查询关键词涉及多个不相关的主题时(例如,查询"美洲虎"至少涉及 4 个不同的主题:(1) 美洲虎汽车;(2) 美洲虎橄榄球队;(3) 美洲虎产品;(4) 哺乳动物美洲虎),构建的查询子图通常会包含若干个不连通的区域.以前的主题提取算法[3,5,8~10]通常都是假定只有一个相关的主题域.在查询范围中如果出现多个主题,它们通常只返回最为流行的主题内容.这显然不能满足许多深入查询的要求.Davison 等人[13]也认识到这个问题,并发现在很多情况下,非主特征向量中会包含一些很有价值的信息.因此,他们试图计算尽可能多的非主特征向量,并通过一组启发策略从这些特征向量中计算网页的权威值和中心值.与他们的方法不同,我们的算法将图中出现的每一个不连通的子图都考虑为一种潜在的主题,通过适当的参数设定,算法可以对每一个主题分别加以处理,并计算出相应的权威值和中心值.计算仍采用迭代的过程,以保持其较高的效率.我们将这种算法称为基于相似度模型的主题发现与提取(similarity-based topic exploration and distillation)算法.

1.7 主题提取技术与文献引用分析(citation analysis)的关联

文献引用分析技术是研究科学文献间相互引用的模式.一种期刊的著名程度通常用"影响因子"来衡量.度量的依据通常是文献被引用的次数.HITS 算法和 PageRank 算法从方法上说都与这种引用分析技术有着紧密的联系.

在 Web 中,一个网页的"影响因子"可以简单地对应为所有指向该网页的链接数目.但是这种简单的度量方法在实际应用中并不合适,它通常会使一些有关广泛主题的网页获得很高的分值(例如 http://www.nytime.com,因为有很高的被链接数目),而不管查询的特定主题.PageRank 算法就存在这样的问题,尽管它并没有这样简单地定义网页的价值度量.事实上,即便在文献分析领域,研究者们也在试图改进这种简单的计算引用次数的度量方法.例如,是否存在一种权重方案以区分不同引用的重要程度呢?显然,这个问题的困难在于重要程度的定义会是一个循环定义.这个问题与 HITS 算法中的权威性(authority)和中心性(hub)定义的情况类似.早在 1976 年,Pinski 和 Narin[14]就通过利用一种迭代的方法,为文献计算出一个重要性分值,他们称其为影响性权重(influence weights),从而克服了这种循环定义的问题.他们的这种方法和后来的 PageRank 算法的思想非常接近.Pinski 和 Narin 所提出的影响性权重方法与 HITS 算法不同的是,Pinski 和 Narin 的方法没有区分权威性(authority)和中心性(hub)这两个概念,权重的影响只是从一个权威性文献直接传播到另一个权威性文献.

HITS 算法与 Pinski 等人的方法的不同点反映出互联网页和传统的引用文献之间的一个基本的区别.在互联网环境中,权威网页之间经常并不存在相互的认可(例如,Netscape 和 Microsoft 之间并不存在任何的链接),因此,只考虑权威性度量的方法在很多情况下无法完成这种权威性的传播过程.HITS 算法通过中心网页的概念解决了这个问题.而在科技文献的引用中,引用关系通常是一种相互交叉的无偏见方式,这使得中心性的概念在度量科技文献的领域并不具备很大的意义.

2 互联网社区的发现技术

正如我们在本文开始部分所指出的,整个互联网存在着成千上万的社区.这些社区有的已经以非常清晰的形式表现出来(例如,门户网站中的层次目录式结构,Yahoo's Recreation: Automotive: Makes and Models: Porsche: Boxster).类似这种通过手工分类展现出的社区到目前为止大约有两万多个[15].但是,在极度分散和无

序的互联网环境中,有理由认为必然还存在更多潜在的未被发现和定义的互联网社区.从互联网中系统地抽取这些社区至少有以下 3 方面的意义:

- 这些社区为了解互联网用户的兴趣提供了有价值的,甚至是最及时、最可靠的信息.
- 这些社区展现了互联网社会学(sociology of web),研究和发现这些社区可以深入了解互联网的进化过程.
- 门户网站通过识别和区分这些社区,可以更有效地组织它们的目录层次(因为很多潜在的社区以很快的速度在增长,而很多已经清晰出现的社区又在逐渐消失).这同时意味着互联网的自动分类成为可能.

这方面的研究也以 IBM Almaden 实验室为代表,在近两年内取得了一些重要的进展[16].其研究的基础仍然起源于上一节所介绍的链接分析算法 HITS.下面,我们将着重介绍两种重要的网上社区定义方法和发现技术.

2.1 基于HITS算法的网上社区发现

研究网上社区的一个首要问题就是要为互联网的链接结构给出一种合理的描述.Kleinberg 等人[10]在提出 HITS 算法之后,进一步用中心性网页和权威性网页这两个基本概念来描述互联网的链接结构特征.他们认为,尽管互联网页的创建和链接过程是一种分散和难以描述的过程,但是,从全局的角度来理解,这些互不关联的创建过程由于作者某种共同的偏好而使得相互间产生了越来越紧密的联系.当这种联系达到一定程度时,某个潜在的社区便产生了.

Gibson 和 Kleinberg 等人[17]在这方面的研究是基于链接分析搜索算法 HITS.HITS 算法模型提供了一种很自然的方式,将链接的结构用一组中心性网页和权威性网页展现出来,尽管这些中心性网页之间并不知道相互的存在,同样,这些权威性网页间也不了解相互的存在,但是可以将这一组中心性网页和权威性网页理解为一个网上的社区.由于这种网上社区的定义是完全基于结构的,所以,可以在不知道特定主题的情况下发现它们.当然,这并不是说将整个互联网分割成许许多多个这样的社区.Kleinberg 等人认为需要通过用一小部分高权威性的相关网页来表示社区的主题.

Gibson 和 Kleinberg 等人的研究主要集中在 3 个问题上.第一,通过 HITS 算法产生的网上社区究竟是什么;第二,用 HITS 算法发现的网上社区是否依赖于根集合的选择;第三,在通过多少次迭代以后,社区的主题才会出现.下面我们主要介绍前两个问题.通过实验和观察,他们发现对于足够广泛的主题,HITS 算法发现社区的效果比较理想,同时,所发现的主题社区也和 HITS 算法根集合的选择几乎无关(即根集合很强的鲁棒性).他们还发现,利用 HITS 算法除了依赖于主题的广泛性之外,同时还依赖于互联网上的知识结构,那些流行的学科(例如,计算机科学)在互联网上表现为存在更多的甚至更合理的链接,因此 HITS 算法具有较好的结果.相反,如果针对那些在互联网中体现较少的学科,HITS 的效果便不令人满意了.他们还发现,对于一些在互联网中存在较少链接的主题,HITS 算法似乎能对主题作某种程度的概念上升和抽象.也就是说,尽管主题并没有明确广义和非广义的界限,但是 HITS 算法倾向于将返回比给定主题更一般的主题(例如,比原主题大但却包含原主题).

2.2 基于二分有向图(bipartite directed graph)的网上社区发现

虽然第 2.1 节的方法给出了一种网上社区的定义,并通过 HITS 算法提炼出社区的主题,但是这种方法仍然缺乏对网上社区的一种清晰描述,从而使得在没有给出特定主题描述时,无法从互联网中有效地抽取潜在的社区(尽管 Gibson 等人也宣称他们所定义的网上社区本身仅依赖于互联网的结构,而不是任何某种预先有计划的构建).

不同于 Gibson 等人的方法[17],Kumar 等人[15]从二分有向图的角度对互联网上的社区给出了一种明确的定义描述.首先考虑如下的图结构 K_{ij},图中的节点分为 F 和 C 两个集合,集合 F 中有 i 个节点,集合 C 中有 j 个节点,并且每个集合 F 中的节点和所有在集合 C 中的节点都存在一条有向边.这样,一种图结构 K_{ij} 被称为完全二分有向图 K_{ij}.根据随机二分图的理论,一个足够大而稠密的随机二分图将以很高的概率包含一个完全二分有向图.那么,如果将某个社区的链接结构看作一个大而稠密的二分有向图,则社区的核就可以用一个完全二分有向图 K_{ij} 来表示.具体到互联网环境中,可以对上述概念有如下直观的理解:如果在互联网上存在一个某种主题的社区,那么这种二分的核必将包含在其中.图 1 给出了一个 K_{43} 二分核的例子.图中左边 4 个节点都存在指向 3 家著名

飞机制造商主页的链接.

接下来的问题就是如何确定核的两个参数,以及采用怎样的方法从整个互联网结构图中枚举出所有社区的核.Kumar 等人并没有为核参数 i,j 声明特定的值.事实上,在实验中他们把 i 的值设在 3~6 之间,而 j 的值设在 3~9 之间.i 和 j 在取值范围内所有的组合均作为实验所要枚举的社区的核.为了适应互联网庞大的规模,Kumar 等人为枚举开发了专门的算法,被称为消去-产生法(elimination-generation).其具体的细节参见文献[15].发现了这些社区的核,那么用类似 CLEVER 系统中的算法就可以发现社区的其余部分.

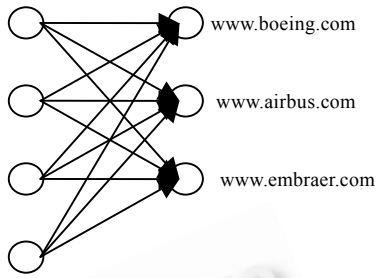

Fig.1 A K_{43} bipartite core
图 1 K_{43} 二分核

因为集中式服务器要处理大量的请求,所以系统对这些服务器的要求很高,集中式服务器还需要维持整个网络中全局名的惟一性.而在 BestPeer 系统中,LIGLO 服务器的数量不受限制,每一个 LIGLO 服务器只需保存它的成员名字并维持成员内名字的惟一性.LIGLO 服务器可以根据自身的能力指定成员的数量,因此,在 BestPeer 系统中不会对 LIGLO 服务器有很高的要求.当 LIGLO 服务器中的成员数达到极限的时候,LIGLO 服务器会拒绝分配新的 BPID,以保证自身的效率.当然,被拒绝的 Peer 可以去另外的 LIGLO 上进行申请.

3 互联网结构图的建模研究

将互联网的整个结构图作为对象来研究不仅对理解互联网的各种属性有直接的意义,同时还对很多互联网算法(例如搜索、爬取以及社区发现等)都有着重要的帮助.另一方面,在研究这些互联网算法的同时,很多实验和观察也进一步促进了互联网图的研究.

对互联网图结构的特征研究首先从链接数目的分布规律开始.在这方面最早的观察[15]证实了链入数目的规律是遵从幂定律的,即具有 i 个链入(in-degree)的网页数正比于 $1/i^x$,x 为大于 1 的数.随后,Albert 等人[18]和 Broder 等人[19]在不同的规模和时间上进一步验证了这个规律.在所有这些实验中,幂定律中指数 x 的取值惊人地一致,都为 2.1.最大规模的实验结果[19]在这里重新给出,如图 2 和图 3 所示.

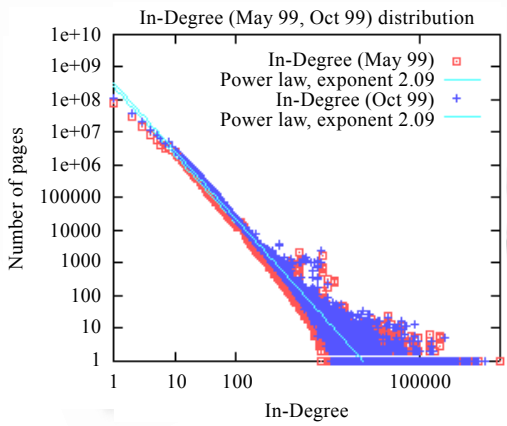

Fig.2 In-Degree distribution of the hyperlinks
图 2 网页的链入数(in-degree)分布

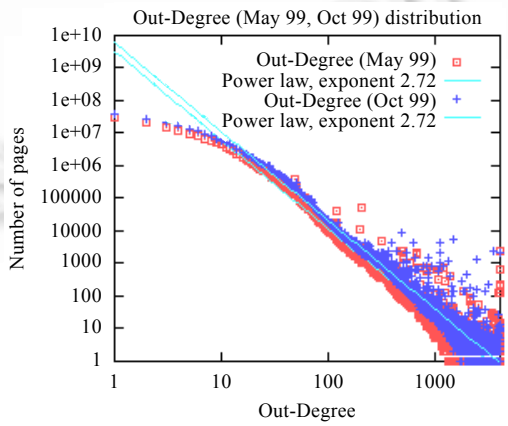

Fig.3 Out-Degree distribution of the hyperlinks
图 3 网页的链出数(out-degree)分布

如果将 Web 看作一个超大的图,那么就有必要对图的连通域加以分析.Broder 等人[19]在基于 AltaVista 1999 年 5 月所爬取的两亿多个网页的分析,为连通域作了如下的分类:

(1) 弱连通域:一组网页的集合,如果它们之间的链接不考虑方向,那么集合中的任何网页均可以通过链接到达任意集合中其他的网页.在 AltaVista 的数据中,最大的弱连通域大约有 1.8 亿个节点,也就是说,至少有 90%的网页是属于这种连通域的.

(2) 强连通域:一组网页集合,其中任意两个节点之间都存在至少一条有向的路径.从互联网的角度来说,也

就是从其中任何一个网页,浏览者可以跟随链接浏览到集合中任何其他网页.实验数据中最大的强连通域包含大约 0.56 亿个节点,次大的强连通域大约有 50 000 个节点,比最大的数量上少 3 个数量级.研究还发现,连通域的规模同样也符合幂定律的分布规律.

在总结和发现了诸如上述的一些互联网特征后,能否用一种很好的数学模型对整个互联网图进行建模呢?很直观地,随机图模型是一种最接近互联网结构图的数学模型.但是,Kumar 等人[20]指出,传统的随机图模型在描述 Web 结构上存在明显的缺陷.其本质原因就是互联网的创建过程虽然是分布的、无计划的过程,但绝不是随机的过程.由于链接包含了人的判断,因此,链接的创建与已有的链接存在某种依赖的关系.

具体来说,如果一个作者关注某个主题,那么他很有可能将现有的一些网页资源列表作为自己所要创建的链接.可以把这种现象称为"拷贝".当然,这并不是说新的作者完全将现有的链接加以物理复制.这里只是认为新作者在现有的主题内创建了链接,而这些链接所指向的网页已经存在一些现存的资源列表上了.另外一点值得注意的是,这个讨论的过程不是偏向于用户模型,而是针对局部的链接创建过程的,通过这些局部过程的积聚,便产生了 Web 的结构和属性.

网上社区的产生过程有助于理解这种从局部到整体的过程.首先,一些用户在互不知道的情况下创建了相同主题的网页,此时这个主题的社区尚未形成.然后,感兴趣的作者开始在主题内链接这些网页,同时创建一些资源列表网页,以帮助有兴趣的第三方找到主题.最终,尽管整个互联网的全局结构仍然是稀疏的,但是局部稠密的子图却伴随着其相关的主题显现出来.

Kumar 等人[20]提出了一种随机拷贝的机制,模拟互联网环境中的链接创建过程.这是一种非常简单而有效的机制.这种机制可以产生链接数目的幂定律分布以及链接创建中的相关性,从一定程度上反映了互联网的真实特点.下面简要介绍基于这种复制机制的互联网图模型.

传统的图模型是静态的,即模型一旦被创建,则图中节点和边的数目就固定了.但是,互联网图模型却要求有新的节点和边随着时间的变化不断在图中出现,而已有的一些节点到后来却要在图中消失.为了描述模型的这些性质,需要用到几个术语.一个模型需要由 4 种随机过程来表征:分别是创建节点和边的过程 C_v,C_e 以及删除节点和边的过程 D_v,D_e.每个随机过程应该是一个由当前时间点和图状态所决定的离散时间过程.

考虑由这种模型产生互联网图的一个例子,C_v 在时刻 t 以独立于当前图的概率 $a_c(t)$ 创建了一个节点.而过程 D_v 以概率 $a_d(t)$ 将一些经过挑选的节点和相关的边从图中删除.这些概率的具体取值可以依据实际互联网增长的速率以及网页生存的半衰期来确定.相对应的边的过程则需要结合随机拷贝的机制.在每个时间点,为每个新加入的节点添加相联系的边,同时还为一些已存在的节点更新它们相关联的边,以模拟互联网结构的更新.对于每个被选中的节点,C_e 过程将随机选择一定数目(例如 k)的边加到该节点上.以某个概率 b,C_e 将这 k 条边随机地加到任意的 k 个目的节点上,而以 $1-b$ 的概率通过拷贝的机制将 k 条边添加到选定的节点上.例如,可以按照某种分布选择一个已有的节点 q,随机地选择 k 条在 q 上的边,将这 k 条边的目的节点作为当前节点创建边的目的节点.如果 q 中包含的链接数少于 k,则再选择另一个节点,重复上述过程,直至拷贝到所需要的数目.同样地,过程 D_e 在时刻 t 以概率 $u(t)$ 在某个分布中选择一条边从图中删除.基于上述模型框架,文献[20]提出了一种简单的互联网图模型的实例.该实例清晰地说明了拷贝机制的确产生了链接数的幂定律分布规律.

4 链接结构在其他方面的应用

除了上述领域,链接结构分析技术在其他超文本信息检索领域[21~25]和网页爬取(crawling)等方面[26~29]也得到了非常有效的应用.本节将对以往这些方面的重要工作作一个简要的回顾.

完全基于文本特征的经典信息检索技术已被深入地进行过研究.由于超文本的链接与文献的引用具有天然的相似性,许多学者从 20 世纪 80 年代中期开始陆续利用链接信息研究超文本的信息检索技术[22,30].虽然这些技术各不相同,但是它们都遵循同一个基本思想来利用链接信息,即从被引用的文档中抽取重要的特征补充到引用的文档中.值得注意的是,从这些技术的基本思想来说,因为并未利用链接的结构信息,所以并不能称其为真正意义上的链接结构分析技术.

直到近些年来,随着互联网的迅速增长,许多学者才开始考虑将互联网的链接拓扑结构作为一种重要的信

息应用到了互联网页的聚类中.文献引用分析(co-citation analysis)[12]的思想成为这方面研究的基本出发点.在第 1.7 节,我们已经深入分析了引用分析与主题提取技术的关联.而在超文本聚类方面,与引用分析的联系则体现得更为直接[31~34].Larson[35]在 1996 年首先提出了直接应用引用分析技术的互联网网页聚类方法.随后还出现了完全基于链接拓扑结构的超文本聚类系统[36].这种算法不同于以往聚类方法的相似度定义,其基本思路是试图从超文本的链接结构中抽取文档的相似性信息.其具体做法是,考虑将两个节点间的独立路径数目作为相似度的度量.同时,许多研究开始致力于将内容信息、用户浏览信息,甚至超文本文档结构信息与链接拓扑结构信息加以整合,提出了各种更为一般化的聚类模型.Mukherjea 等人[37]提出了一种同时利用结构和内容的交互式超文本聚类算法.在他们的模型中,用户可以精确地描述他们的信息需求,而所有的节点都包含一些内容及其子图结构的信息.但是,这种聚类的算法是半自动的.Weiss 等人[32]和 Modha 等人[38]分别提出了两种将文档内容和链接结构整合在一起的全自动聚类算法.这两种算法的基本出发点比较接近,都是将文本信息和结构信息分别表示为独立的向量,再通过算法模型对它们加以整合.但在算法模型的选择、相似性的度量以及聚类的表示上,上述两种算法仍存在较大的差别.Pirolli 等人[39]不仅考虑了文档的内容信息和链接的结构信息,而且还将用户的浏览信息整合到聚类算法中,不同于文献[38],文献[39]将超文本的文本特征、链接特征以及用户的浏览特征通过一个单一的向量来加以表示.Chen[40]还提出了一种一般化的相似度分析方法,用以整合超文本的内容信息、链接特征以及浏览模式.

在网页爬取研究方面,链接结构信息的利用是一种必需的手段.传统的搜索引擎技术将海量的互联网文档存储在本地,并提供一种方法来对它们加以查询.这样的策略将面临可伸缩性(scalability)问题、存储(storage)问题,甚至性能(performance)上的问题,并且这些问题还将随着互联网规模的继续增大而更加突出.因此,依据特定要求爬取网页的方法成为解决这一问题的一种有效途径.

智能爬取网页的研究历史也不长,其中 Chakrabarti 等人[28]提出的主题爬取(focused crawling)方法最具代表性.该方法的基本出发点就是认为在 Web 图上存在小范围的主题区域,并可以通过一些技术选择一些好的进入点,将爬取的范围维持在这个主题区域内.另外,hub 和 authority 网页[3]的概念也被应用到主题爬取的技术中.Chakrabarti 等人认为,相关主题的资源将以 hub 网页(包含大量同一主题链接的网页)或者 authority 网页(其内容对应于相关主题的网页)的形式出现.利用 hub-authority 模型,主题爬取方法的分类概率模型能够成功地选择与主题相关的网页.实验结果表明,这是一种有效的网页爬取方法.

Aggarwal[41]等人认为,尽管主题爬取的方法已经非常有效,但是这种方法过于依赖用户给定信息的质量.而实际上仍然有很多信息是算法可以主动利用的,诸如已爬网页的文本信息、待爬网页的 URL 信息等.他们提出了一种更为一般化的网页爬取框架.这种方法满足了更为任意的用户查询,包括主题查询、关键词查询甚至是二者的结合.

这方面还有一些其他的研究工作着眼于辅助现有的搜索引擎技术:有的着重于搜索引擎辅助缓存方面的研究,例如文献[42,43];有的则关注对现有固定互联网数据库有效的更新,这类技术通常具备特定的功能(例如数据挖掘[27,44,45]).尽管应用角度不同,这些技术从网页爬取研究的角度而言,仍然只是将互联网页的一个子集作为研究的对象.

5 未来的研究方向

在本文中,我们力图详尽地回顾近些年来链接结构在互联网领域中各个方面的研究和应用.链接分析作为一种研究超文本环境极为重要的工具,在互联网研究领域具有极其重要的理论研究价值和广泛的应用背景.同时,链接分析的研究从总体上说尚处于一个起步的阶段,已有的研究工作正为这个领域提出越来越多需要解决的问题.下面是我们正在从事的和一些可能的研究方向.

(1) 文献[4]中的主题提取模型给出了一种考虑更为一般化的相似度定义.但是,由于这种一般相似度所对应的图不再是常规意义上的图,而是一个超图,所以文献[4]考虑了一种折衷的方案,即通过一个映射将这种一般性的相似度还原为两两相似度.但是,应用超图的遍历将有可能保留这种一般相似度的定义而直接完成主题提取的过程,从而避免了映射所带来的信息缺损.

(2) 基于 HITS 算法的 ARC 和 CLEVER 系统[5,9]尽管考虑了网页文本的内容,提出了改良的迭代算法,但是这种算法在针对狭义的查询主题时仍然不能取得另人满意的结果.因此,文本信息的结合方法仍然可能存在更有效的方式.例如,通过区分不同网页的创建风格,将文本信息整合到权重中仍然存在很多其他可能的方式.

(3) 基于链接分析的方法无一例外地遵从如下的假设:每个链接代表一个网页作者对所指向的网页的一种独立的"认可".但是,在真实的互联网环境中存在大量重复的链接、镜向的站点以及商业广告链接,这些形成了很多裙带的链接关系,从而使上述假设不再成立.事实上,任何简单的方法都无法区分这种带有偏向性的,甚至是无意义的裙带链接关系.深入研究并区分这种裙带链接对链接分析技术更为有效的应用具有十分重要的意义.

(4) 虽然在互联网社区方面,文献[15,17]已做了一些探索,但是应该认识到现有的这些定义仍然是比较初步的.网上社区虽然是由互联网的链接结构所决定,但它本身具备概念的特点.如何从语义上组织这种网上社区,是一个值得研究的课题.

(5) 用二分图定义网上社区[15]并不是一种惟一的方式.基于统计学习基础的网上社区的数学描述和定义可能是另一个非常有希望的努力方向.

(6) 网页的超链接和浏览之间存在着密不可分的联系.互联网链接结构很大程度上决定了人们浏览互联网的过程.有理由相信,在清楚互联网社区的前提下,更为有效地利用链接信息加强现有的搜索算法和浏览工具将会成为可能.

(7) 依据文献[20]所提出的互联网图模型究竟会产生一种怎样的随机图属性和进化过程?与实际互联网的属性和进化过程是否一致?这方面的实验研究是十分必要的.

互联网图连通域的分类是否只能被分为强连通域和弱连通域[19],是否还存在其他有价值的分类概念?互联网图的衍生图(例如通过共同引用关系所导出的无向图)将会存在怎样的结构,应该用什么样的概念来描述这种结构?这些都是值得进一步深入研究的问题.

References:

[1] Chakrabarti S, Dom B, Kumar S, Raghavan P, Rajagopalan S, Tomkins A, Gibson D, Kleinberg S. Mining the Web's link structure. IEEE Computer,1999,32(8):60~67.

[2] Chakrabarti S. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In: Shen VY, Saito N, Lyu MR, Zurko ME, eds. Proceedings of the 10th International World Wide Web Conference. Hong Kong: ACM Press, 2001. 211~220.

[3] Bharat K, Henzinger M. Improved algorithms for topic distillation in a hyperlinked environment. In: Voorhees E, *et al*., eds. Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval. Melbourne: ACM Press, 1998. 104~111.

[4] Wang X, Wu H, Wei L, Zhou A. A similarity-based analysis model for topic distillation. International Journal of Computational Intelligence and Application, 2002,2(3):267~275.

[5] Chakrabarti S, Dom B, Gibson D, Kleinberg J, Raghavan P, Rajagopalan S. Automatic resource compilation by analyzing hyperlink structure and associated text. In: Thistlewaite P, *et al*., eds. Proceedings of the 7th ACM-WWW International Conference. Brisbane: ACM Press, 1998. 65~74.

[6] Dean J, Henzinger M. Finding related pages in the World Wide Web. Computer Networks, 1999,31(11-16):1467~1479.

[7] Borodin A, Roberts G, Rosenthal J, Tsaparas P. Finding authorities and hubs from link structures on the World Wide Web. In: Shen VY, Saito N, Lyu MR, Zurko ME, eds. Proceedings of the 10th ACM-WWW International Conference. Hong Kong: ACM Press, 2001. 415~429.

[8] Chakrabarti S, Dom B, Gibson D, Kumar S, Raghavan P, Rajagopalan S, Tomkins A. Experiments in topic distillation. In: Proceedings of the ACM SIGIR workshop on Hypertext Information Retrieval on the Web. Melbourne: ACM Press, 1998. 13~21.

[9] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. In: Thistlewaite P, *et al*., eds. Proceedings of the 7th ACM-WWW International Conference. Brisbane: ACM Press, 1998. 107~117.

[10] Kleinberg J. Authoritative sources in a hyperlinked environment. In: Tarjan RE, *et al*., eds. Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms. New Orleans: ACM Press, 1997. 668~677.

[11] Bharat K, Broder A, Henzinger M, Kumar P, Venkatasubramanian S. The connectivity Server: Fast access to linkage information on the Web. In: Thistlewaite P, *et al*., eds. Proceedings of the 7th ACM-WWW International Conference. Brisbane: ACM Press, 1998. 469~477.

[12] White H, McCain K. Visualizing a discipline: An author co-citation analysis of information science 1972−1995. Journal of the American Society for Information Science, 1998,49(4):327~356.

[13] Davison B, Gerasoulis A, Kleisouris K, Lu Y, Seo H, Wang W, Wu B. DiscoWeb: Applying link analysis to web search (extended abstract). In: Albert V, *et al*., eds. Poster Proceedings of the 8th ACM-WWW International Conference. Toronto: ACM Press, 1999. 148~149.

[14] Pinski G, Narin F. Citation influence for journal aggregates of scientific publications: Theory with application to the literature of physics. Information Processing & Management, 1976,12:297~312.

[15] Kumar S, Raghavan P, Rajagopalan S, Tomkins A. Trawling emerging cyber-communities automatically. In: Albert V, *et al*., eds. Proceedings of the 8th ACM-WWW International Conference. Toronto: ACM Press, 1999. 1481~1493.

[16] Chakrabarti S. Recent results in automatic Web resource discovery. ACM Computing Survey, 1999,31(4):21~27.

[17] Gibson D, Kleinberg J, Raghavan P. Inferring Web communities from link topology. In: Proceedings of the 9th ACM Conference on Hypertext and Hypermedia. Pittsburgh: ACM Press, 1998. 225~234.

[18] Albert R, Jeong H, Barabasi A. Diameter of the World Wide Web. Nature, 1999,401:130~133.

[19] Broder A, Kumar R, Maghou F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J. Graph structure in the Web. In: Albert V, *et al*., eds. Proceedings of the 9th ACM-WWW International Conference. Amsterdam: ACM Press, 2000. 309~320.

[20] Kumar R, Raghavan P, Rajagopalan S, Sivakumar D, Tomkins A, Upfal E. The Web as a graph. In: Serge A, ed. Proceedings of the 18th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Pennsylvania: ACM Press, 1999. 109~118.

[21] Carriere J, Kazman R. WebQuery: Searching and visualizing the Web through connectivity. Computer Networks and ISDN Systems, 1997,29(8-13):1257~1267.

[22] Chakrabarti S, Dom B, Indyk P. Enhanced hypertext classification using hyperlinks. In: Laura H, ed. Proceedings of the ACM SIGMOD International Conference on Management of Data. Washington: ACM Press, 1998. 307~318.

[23] Spertus E. ParaSite: Mining structural information on the Web. Computer Networks and ISDN Systems, 1997,29(8-13):1205~1215.

[24] Cooley R, Mobasher B, Srivastava J. Web mining: Information and pattern discovery on the World Wide Web. In: Enrico P, ed. Proceedings of the 9th International Conference on Tools with Artificial Intelligence. Newport Beach: IEEE Computer Society, 1997. 558~567.

[25] Chen C, Czerwinski M. From latent semantics to spatial hypertext—An integrated approach. In: Proceedings of the 9th ACM Conference on Hypertext and Hypermedia. Pittsburgh: ACM Press, 1998. 77~86.

[26] Diligenti M. Focused crawling using context graphs. In: Amr A, *et al*., eds. Proceedings of the 26th International Conference on Very Large Data Bases. Cairo: Morgan Kaufmann Publishers, 2000. 527~534.

[27] Cho J, Garcia-Molina H, Page L. Efficient crawling through URL ordering. In: Thistlewaite P, *et al*., eds. Proceedings of the 7th ACM-WWW International Conference. Brisbane: ACM Press, 1998. 161~172.

[28] Chakrabarti S, van den Berg M, Dom B. Focused crawling: A new approach to topic specific resource discovery. In: Albert V, *et al*., eds. Proceedings of the 8th ACM-WWW International Conference. Toronto: ACM Press, 1999. 1623~1640.

[29] Najork M, Wiener J. Breadth-First search crawling yields high-quality pages. In: Shen VY, Saito N, Lyu MR, Zurko ME, eds. Proceedings of the 10th International World Wide Web Conference. Hong Kong: ACM Press, 2001. 295~308

[30] Croft B, Turtle H. A retrieval model for incorporating hypertext links. In: Proceedings of the 1st ACM Conference on Hypertext and Hypermedia. Pittsburgh: ACM Press, 1989. 213~224.

[31] Pitkow J, Pirolli P. Life, death, and lawfulness on the electronic frontier. In: Steven P, ed. Human Factors in Computing Systems, CHI'97 Conference Proceedings. Atlanta: ACM Press, 1997. 3~10.

[32] Weiss R, Velez B, Sheldon M. Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In: Proceedings of the 7th ACM Conference on Hypertext. Washington DC: ACM Press, 1996. 180~193.

[33] Ding J, Gravano L, Shivakumar N. Computing geographical scopes of Web resources. In: Amr A, *et al*., eds. Proceedings of the 26th International Conference on Very Large Data Bases. Cairo: Morgan Kaufmann Publishers, 2000. 545~556.

[34] Bar-Yossef Z. Approximating aggregate queries about Web pages via random walks. In: Amr A, *et al*., eds. Proceedings of the 26th International Conference on Very Large Data Bases. Cairo: Morgan Kaufmann Publishers, 2000. 535~544.

[35] Larson R. Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. In: Hans-Peter F, *et al*., eds. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Zurich: ACM Press, 1996. 85~92.

[36] Botafago A. Cluster analysis for hypertext systems. In: Robert K, *et al*., eds. Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval. Pittsburgh: ACM Press, 1993. 116~125.

[37] Mukherjea S. WTMS: A system for collecting and analyzing topic-specified web information. In: Albert V, *et al*., eds. Proceedings of the 9th ACM-WWW International Conference. Amsterdam: ACM Press, 2000. 457~471.

[38] Modha D, Spangler W. Clustering hypertext with application to web searching. In: Proceedings of the 11th ACM Conference on Hypertext and Hypermedia. San Antonio: ACM Press, 2000. 143~152.

[39] Pirolli P, Pitkow J, Rao R. Silk from sow's ear: Extracting useable structures from the Web. In: Proceedings of the CHI'96 Conference on Human Factors in Computing Systems. Vancouver: ACM Press, 1996. 118~125.

[40] Chen C. Structuring and visualizing the WWW by generalized similarity analysis. In: Mark B, *et al*., eds. Proceedings of the 8th ACM Conference on Hypertext. Southampton: ACM Press, 1997. 177~186.

[41] Aggarwal C, Al-Garawi F, Yu P. Intelligent crawling on the World Wide Web with arbitrary predicates. In: Michael RL, ed. Proceedings of the Tenth International World Wide Web Conference. Hong Kong: ACM Press, 2001. 96~105.

[42] Wills C, Mikhailov M. Towards a better understanding of web resources and sever responses for improved caching. In: Albert V, *et al*., eds. Poster Proceedings of the 8th ACM-WWW International Conference. Toronto: ACM Press, 1999. 1231~1243.

[43] Douglis F, Feldmann A, Krishnamurthy B. Rate of change and other metrics: A live study of World Wide Web. In: Proceedings of the 1st USENIX Symposium on Internet Technologies and Systems. Monterey: USENIX Press, 1997. 87~101.

[44] Cho J, Garcia-Molina H. The evolution of the web and implications for an incremental crawler. In: Amr A, *et al*., eds. Proceedings of the 26th International Conference on Very Large Data Bases. Cairo: Morgan Kaufmann Publishers, 2000. 200~209.

[45] Cho J, Garcia-Molina H. Synchronizing a database to improve freshness. In: Weidong C, *et al*., eds. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Dallas: ACM Press, 2000. 117~128.