PH-MaxFlow 算法发现 Web 社区

郭希娟[1],刘 静[2]
GUO Xi-juan[1],LIU Jing[2]

燕山大学 信息科学与工程学院,河北 秦皇岛 066004
College of Information Science and Engineering,Yanshan University,Qinhuangdao,Hebei 066004,China
E-mail:l6616j@yahoo.com.cn

GUO Xi-juan,LIU Jing.Mining Web communities with PH-MaxFlow algorithm.Computer Engineering and Applications, 2009,45(5):113-116.

Abstract: HITS is a classical algorithm for the computation of the authority value and hub value of Web pages using link technology,it can mine Web communities related to some topic quickly,but sometimes there is "Topic Drift" phenomenon.This paper presents PHITS algorithm controlling the"Topic Drift" phenomenon to a certain extent,the PH-MaxFlow algorithm using the pages with high authority value as seeds can mine precise communities.An effective method is presented to appraise the identified Web communities.The results of experiment show that the PH-MaxFlow algorithm can mine more reasonable Web communities.
Key words: Web communities;Hyperlink-Induced Topic Search(HITS) algorithm;maximum flow algorithm

摘 要:HITS 是一种经典的利用链接技术计算网页权威值和中心值的算法,它能够快速发现主题相关网页,其缺点是会发生"主题偏移"现象,首先提出 PHITS 算法,在一定程度上抑制了这种现象的发生。运用该方法提取权威值高的页面,作为 PH-MaxFlow 算法的种子节点,使得发现的 Web 社区更精确。同时提出了一种有效的评价 Web 社区的标准,用这个标准对原始最大流算法和提出的 PH-MaxFlow 算法进行比较,从而得出 PH-MaxFlow 算法发现的 Web 社区与主题更相关。
关键词:Web 社区;基于超链接分析的主题搜索算法;最大流算法
DOI:10.3778/j.issn.1002-8331.2009.05.033 **文章编号**:1002-8331(2009)05-0113-04 **文献标识码**:A **中图分类号**:TP393

1 引言

随着互联网的广泛应用,互联网已经成为一个巨大的、分布广泛的全球信息服务中心。可以将互联网抽象成一个庞大的 Web 图,这个图由很多的点和边组成,这些点和边都包含着丰富的互联网的信息,其中点代表互联网上的页面,边代表页面之间的超链接。

Web 社区可以松散地定义为基于某个特定主题下的、相互连接的 Web 页面集,且社区内页面的链接密度大于社区外页面的链接密度[1]。通常上讲一个 Web 社区只是整个互联网 Web 图中的一个非常小的子图,如何去发现互联网上这些潜在的 Web 社团进而发现潜在的互联网信息也是近几年众多研究者关注的研究领域,提出了各种各样的基于链接结构分析发现 Web 社区的方法。Gibson 和 Kleinberg 等人[2]提出了基于链接分析的搜索算法 HITS,Kumar 等人[3]从二分有向图的角度对 Web 社区给出了一个明确的定义描述,把 Web 社区看作是一些二分有向图的核,Asano 等人[4]提出了面向站点而不是页面的方法,Flake 等人[5]最先提出通过最大流算法来发现 Web 社区,文献[6]对基于二分有向图算法和基于最大流算法进行了比较,文献[7]将文献[4]与最大流算法相结合阐述了运用站点信息的优势,文献[9]将 Web 社区的发现与定义稠密连接子图相联系,主要运用于在很大的 Web 图中发现 Web 社区。针对原始最大流算法种子节点发现的环节,用 PHITS 算法发现种子节点,使发现种子节点的过程与用 PH-MaxFlow 算法发现 Web 社区的过程相分离,用与主题关系更密切的种子节点发现更符合需要的 Web 社区。

2 相关研究工作

2.1 HITS 算法

在链接环境下整个互联网可以看作是一个有向图结构 $G(V,E)$,网页对应图中的节点集 V,网页之间的超链接对应图中的边集 E,如果有链接从 p 指向 q,则 $(p,q)\in E$。

根据 Kleinberg[2]中,Web 页面可分为两种类型,即中心页面(Hub)和权威页面(Authority)。权威页面是指人们公认的在某一主题上内容权威的页面;中心页面是指页面上有很多指向权威页面链接的页面。中心页面与权威页面形成一个相互加强的关系:好的中心页面指向许多好的权威页面,而好的权威页面被许多好的中心页面所指向。

HITS 首先跟据查询主题确定一个网络子图 $G(V,E)$,然后迭代计算出每一个网页的权威值和中心值,具体步骤可分为三步:

作者简介:郭希娟(1959-),女,教授,硕士生导师,主要研究方向:计算机器人技术;刘静(1983-),女,硕士研究生,主要研究方向:Web 社区挖掘。
收稿日期:2008-01-11 **修回日期**:2008-04-21

(1)通过文本分析进行关键字匹配得到与主题最相关的 K 个网页(K=200)的集合,称之为 root 集。

(2)通过链接分析扩展 root 集,扩展后得到的集合称之为 base 集。扩展方法是:对于 root 集中任一网页 p,加入所有 p 中链接所指向的网页到 root 集,加入最多 d(d=50)个指向 p 的页面到 base 集。

(3)计算 base 集中所有页面的中心度和权威度:若 G 中有 n 个节点,设 n 维向量 a、h,$a^{(p)}$、$h^{(p)}$ 分别表示节点 p 的权威值和中心值。规定以下两个操作:

$$a^{(p)}=\sum_{q:(q,p)\in E}h^{(q)} \tag{1}$$

$$h^{(p)}=\sum_{q:(p,q)\in E}a^{(q)} \tag{2}$$

每一次操作后对 $a^{(p)}$ 和 $h^{(p)}$ 进行归一化,经过若干次的迭代后,可得出每一个页面的权威值和中心值。

2.2 最大流算法

给定一个流网络图 $G(V,E)$,每条边的容量为 $c(u,v)\in Z^{+}$,两个节点 $s,t\in V$,节点 s 为源点(source),节点 t 为汇点(sink)。直观上讲,假设边为管道,节点为开关,那么最大流问题就是如何让源点 s 到汇点 t 能流过的流量最大。

Flake 等人[6]给出的 Web 社区的定义如下:Web 社区是图 $G(V,E)$的顶点集 $C\subseteq V$,C 中所有节点 v 和 C 中其他点连接的边数不小于同 $V-C$ 中节点的连接数。即 $\forall u\in C$,它满足条件 $\sum_{v\in C}w_{uv}\geqslant\sum_{v\in V-C}w_{uv}$,称之为 FLG 社区。也可以表达为:一个 Web 社区就是 Web 网页的集合,满足社区内任意网页同社区内其它网页之间的链接数不小于同社区外网页的链接数。Flake 等人证明了通过执行 s-t 最大流/最小割方法后,可以从始点经非饱和边到达的点都满足 FLG 社区的定义,其中非饱和边指的是边上的流小于边容量的边。通过该方法发现的社区通常比较大,而且出现的孤立点不多,是一种比较有效的社区挖掘方法。

FLG 社区存在边界模糊问题,尤其是在稠密的连通子图中,许多稍微不同的子图都可能是 FLG 社区。这样在给定种子节点的情况下,不能保证所得社区的唯一性,对图的划分也不够准确。

针对此问题,文献[8]提出了更严格的 Web 社区的定义,社区内的点要满足 $\sum_{v\in C}w_{uv}>\sum_{v\in V-C}w_{uv}$,$\forall u\in C$,社区外的点要满足 $\sum_{v\in C}w_{uv}\leqslant\sum_{v\in V-C}w_{uv}$,$\forall u\in V-C$,由此解决了 Flake 提出的原始的最大流算法的边界模糊问题。

3 PH-MaxFlow 算法的提出与实现

在 Flake 等人提出的发现 Web 社区的方法中,算法最初的种子节点的选取是人为的,也就是凭借人的主观意识选取与查询相关的网页作为种子节点,然后反复检查上一次所发现社区中的非种子节点,将非种子节点中链接入度和出度最大的网页节点作为种子节点加入到原始的种子节点集当中,作为下一次算法循环的种子节点,反复执行,直到发现的社区趋于稳定的大小为止。

该算法中发现种子节点存在以下两点问题:

(1)对于新种子节点的选择不好控制,如果上一次所发现社区中的所有非种子节点的链接入度和出度都很低,那么再将之作为新的种子节点加入就使得种子节点的发现变得没有意义。

(2)原始的最大流算法认为入度高的网页和出度高的网页都包含了指向重要网页的链接,所以既选取入度高的节点又选取出度高的节点作为新加入的种子节点。但是在实际查询的过程中,即使用户打开了这种出度高的网页(即中心页面),也很难在该页面中找到指向相关权威页面的链接。此外由于中心页面通常包含很多不相关的链接,如果把出度高的网页作为种子节点,增加了噪音页面被提取出来的可能性。

针对上述两个问题,在发现 Web 社区之前先根据主题需要,用 PHITS 算法发现权威值较高的节点,将之作为 Web 社区发现算法的种子节点,从每个种子节点出发构造 Web 图,在图上运用最大流最小割算法,从而发现 Web 社区,最后再将每个种子节点发现的 Web 社区求并集,即得出最终的社区,将整个 Web 社区发现算法称之为 PH-MaxFlow(PageRankHITS-MaxFlow)算法。

3.1 PHITS 算法发现种子节点

原始的 HITS 算法纯粹基于链接分析来发现权威网页和中心网页,运用此算法时内存中只需存储节点和链接信息,这样在一台 PC 机上只需很少的内存即可实现,并且能够快速发现主题相关网页,但 HITS 算法存在以下三个问题。

(1)单纯分析网页之间的链接信息,不考虑页面本身的重要性,使得分析结果有一定的偏差。

(2)大量的实验表明,拥有同一个主机名、域名和 IP 地址的不同页面通常共用一个服务器,同一个服务器下页面之间的链接通常是为了提高用户访问的方便性,对计算页面权威值没有实际意义。

(3)如果有人故意在一个服务器下制造了指向另一个服务器下同一个页面的很多链接,则必然导致这个页面的权威值不合理地上升,将这种情况称之为“相互加强”现象。

以上三个问题导致“主题偏移”现象的发生,如图 1 所示。为了有效抑制“主题偏移”现象,针对 HITS 算法的第(2)步进行改进,去除同一个服务器下的页面之间的链接,重新构造边集,得到 E';并且在计算权威值时不仅考虑页面之间的超链接,而且考虑页面本身的重要性,即 PageRank 值,页面 p 的 PageRank 值用 $RP^{(p)}$ 表示,对 HITS 算法的第(3)步进行改进,计算页面 p 的权威值和中心值的公式如下:

$$a^{(p)}=\frac{1}{|\{q|(q,p)\in E'\}|}\sum_{q:(q,p)\in E'}PR^{(q)}h^{(q)} \tag{3}$$

$$h^{(p)}=\frac{1}{|\{q|(p,q)\in E'\}|}\sum_{q:(p,q)\in E'}PR^{(q)}a^{(q)} \tag{4}$$

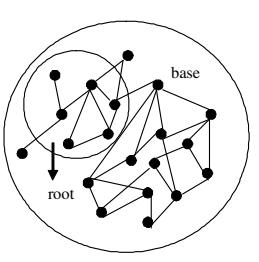

图 1 HITS 算法中的主题偏移

将改进后的算法称之为 PHITS(PageRankHITS)算法。经过一定次数的迭代,当 $a^{(p)}$ 趋于稳定时将 $a^{(p)}$ 值较高的页面节点提取出来,作为 PH-MaxFlow 算法的种子节点。由公式(3)的分母部分可以看出,即使一台服务器上有很多页面指向另一台服务器的同一页面,也不会导致这个页面的权威值不合理地上升,从而避免"相互加强"现象。本文只考虑节点的权威值,对中心值不予考虑,避免将噪音页面提取出来,所以在 PHITS 算法中只计算页面的权威值即可。

3.2 PH-MaxFlow 算法发现 Web 社区

定义 1[8] Web 社区一个 Web 社区是点集 $C\subset V$,满足以下两个条件:

$$\sum_{v\in C} w_{uv} > \sum_{v\in V-C} w_{uv},\ \forall u\in C \tag{5}$$

$$\sum_{v\in C} w_{uv} \leqslant \sum_{v\in V-C} w_{uv},\ \forall u\in V-C \tag{6}$$

根据上面提出的 Web 社区定义,设计 Web 社区挖掘算法应该寻找满足(5)(6)条件的顶点集。由于对于任意 $u\in V_c-\{s\}$,$\sum_{v\in C} w_{uv} > \sum_{v\in V-C} w_{uv}$,且对于任意 $u\in V-V_c-\{t\}$,$\sum_{v\in C} w_{uv} \leqslant \sum_{v\in V-C} w_{uv}$[8],所以要满足条件(5),(6),只需让 s 满足 $\sum_{v\in C} w_{sv} > \sum_{v\in V-C} w_{sv}$,且 t 满足 $\sum_{v\in C} w_{tv} \leqslant \sum_{v\in V-C} w_{tv}$。

定理 1[8] 如果 C_i 和 C_j 都是 Web 社区,那么 $C_i\cup C_j$ 也是Web 社区。

PH-MaxFlow 算法基于定义 1 和定理 1,运用 PHITS 算法发现的权威值较高的节点作为种子节点,给出 Web 社区发现的 PH-MaxFlow 算法如下:

输入:种子节点集 $S\{s_1,s_2,s_3\cdots s_n\}$

输出:Web 社区集合

(1) for S 中的每个节点 s
(2)　从 s 出发,作深度为 2 的爬取,得到图 $G(V,E)$
(3)　设 t 作为虚拟的汇点
(4)　$V=V\cup\{t\}$
(5)　for each $(u,v)\in E$ do
(6)　　$c(u,v)=2$
(7)　　if $(u,v)\notin E$ then
(8)　　$c(v,u)=c(u,v)$
　　end if
　end for
(9)　for $v\in V-\{s\}$
(10)　增加(v,t)到 E,$c(v,t)=1$
　end for
(11)　构建为一个流网络 $N(s,t,V,E,c)$
(12)　使用最大流最小割算法,求得 $C(V_s,V_t)$
(13)　if $\sum_{v\in V_s} w_{sv} > \sum_{v\in V-V_s} w_{sv}$ and $\sum_{v\in V_s} w_{tv} \leqslant \sum_{v\in V-V_s} w_{tv}$ then
(14)　$V_c=V_s$
(15)　else V_c=null
　end for
　返回所有 V_c 的并集$\cup V_c$

图 2 演示了该算法的运作流程。第(2)步从 S 中节点 s 出发,使用深度搜索的方法进行深度为 2 的爬取,构造出图 $G(V,E)$,第(3),(4)步加入一个虚拟的汇点到 V,(5)~(11)步把有向图转换为无向图,并且为每条边都设置容量,构造出一个流网络 $N(s,t,V,E,c)$,(12)~(16)步对该流网络使用最大流最小割算法,找到同 s 通过非饱和边相连的节点集 V_s,如果 $\sum_{v\in V_s} w_{sv} > \sum_{v\in V-V_s} w_{sv}$ 并且 $\sum_{v\in V_s} w_{tv} \leqslant \sum_{v\in V-V_s} w_{tv}$,则 $V_c=V_s$。根据定理 1,返回的 V_c 的并集$\cup V_c$ 就是要寻找的 Web 社区。

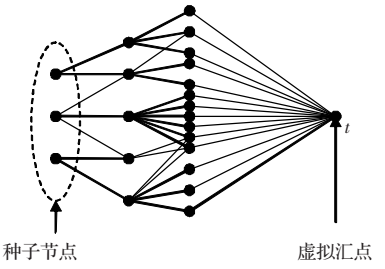

图 2　挖掘算法的图示

4　对所发现的 Web 社区进行评价

判断所发现的社区是否合理,可以用相关度来衡量。设定相关度=社区内同主题相关的网页数/社区总网页数,这是受一定的人为因素所影响,因为可能在某人看来与主题相关的页面在另一个人看来并不相关。为了客观地对结果进行评价,特提出一个评价公式。

根据 PH-MaxFlow 算法,发现了 n 个 Web 社区,构造一个 $n\times n$ 的对称矩阵 $\boldsymbol{B}$,$\boldsymbol{B}$ 的每一个元素 B_{ij} 代表社区 i 和社区 j 之间的连接数占$\cup V_c$ 中链接数的比例,B_{ii} 代表社区内的链接数占$\cup V_c$ 中链接数的比例,矩阵 $\boldsymbol{B}$ 的迹 $tr\boldsymbol{B}=\sum_i B_{ii}$ 给出了在同一个社区中的链接数的总和占$\cup V_c$ 中链接数的比例。好的社区划分,必然拥有较高的 $tr\boldsymbol{B}$ 值。然而 $tr\boldsymbol{B}$ 本身并不能作为衡量划分社区好坏的标准,因为若对某个图不进行划分,且看作是一个社区,则 $tr\boldsymbol{B}$ 的值就必然是 1,这并不能体现出好的 Web 社区的结构。所以,再进一步定义每一行(或列)的元素之和,$B_i=\sum_j B_{ij}$,B_i 代表和社区 i 有链接的链接数占$\cup V_c$ 中链接数的比例。广义讲,一条链接开始于社区 i 并且结束于社区 i,也就是社区内部的链接占$\cup V_c$ 中链接数的比例的期望值是 B_i*B_i,所以可以定义 $\boldsymbol{Q}=\sum_i (B_{ii}-B_iB_i)=tr\boldsymbol{B}-\|\boldsymbol{B}^2\|$ 作为衡量 Web 社区划分好坏的标准,其中 $\|\boldsymbol{B}^2\|$ 代表矩阵 $\boldsymbol{B}^2$ 的所有元素之和。

5　实验设计和结果分析

实验选用 Internet Archive 和 Web crawler 为主题测试提出的 PH-MaxFlow 算法的有效性。

首先使用 Internet Archive 和 Web crawler 为关键词来构造数据集,用 Google 搜索引擎索引 200 个网页构成 root 集,然后把 root 集进一步扩展为 base 集,它包含了所有由 root 集中

的页面所指向的页面,以及所有指向 root 集的页面,把这些网页限制在 50 个以内。由 3.1 提出 PHIST 算法得到 $a^{(p)}$ 值较高的网页为种子节点,再经过 3.2 PH-MaxFlow 算法处理得到所求的社区。用所提出的评价标准对发现的社区进行评价,得到原始最大流算法和提出的 PH-MaxFlow 算法的对比结果,如表 1 和表 2 所示。对比结果,可以看出,使用提出的 PH-MaxFlow 算法,可以发现更权威的种子节点,从而发现更大更精确的社区,并且社区的划分也比原始的最大流算法合理。

表 1 原始最大流算法发现 Web 设区

社区主题	种子网页	社区大小	Q 值
Internet Archive	archivists.org intermemory.org trec.nist.gov	289	0.43
Web crawler	webcrawler.com robotstxt.org/wc/robots.htm	153	0.38

表 2 PH-MaxFlow 算法发现 Web 社区

社区主题	种子网页	社区大小	Q 值
Internet Archive	archive.org marxists.org archivists.org archive.org/web/web.php	326	0.56
Web crawler	webcrawler.com metacrawler.com webcrawler.net.cn	181	0.45

6 结论

首先提出 PHITS 算法,在一定程度上抑制了原始的 HITS 算法所存在的"主题偏移"现象,运用 PHITS 算法发现权威值高的页面,即与主题关系更为密切的页面作为最大流算法的种子节点。采用提出的 PH-MaxFlow 算法,将种子节点的发现过程和 Web 社区的发现过程相分离,避免了原始最大流算法反复检验已发现 Web 社区中非种子节点的入度和出度的繁琐。为了比较原始的最大流算法和提出的 PH-MaxFlow 算法,特提出了客观的评价标准,更具有说服力。实验结果表明运用所提出的算法可以发现更符合主题需要的 Web 社区。

参考文献:

[1] 高琰,古士文,唐琎.基于链接分析的 Web 社区发现技术的研究[J].计算机应用研究,2006(7):183-185.

[2] Kleinberg J.Authoritative sources in a hyperlinked environment[C]//Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms.San Francisco,USA,Jan 25-27 1998.USA,SIAM,Philadelphia,1998:668-677.

[3] Kumar R,Raghavan P,Rajagopalan S,et a1.Trawling the Web for emerging cyber communities[C]//Proc of the 8th WWW International Conference,Toronto,May 11-May 14 1999.Netherlands,Elsevier Sci B V,1999:1481-1493.

[4] Asano Y,Imai H,Toyoda M,et al.Finding neighbor communities in the Web using an inter-site graph[J].IEICE Transactions on Information and Systems,2004,9:2163-2170.

[5] Flake G W,Lawrence S,Giles C L,et al.Self organization of the web and identification of communities[J].IEEE Computer,2002,35(3):66-71.

[6] Imafuji N,Kitsuregawa M.Effects of maximum flow algorithm on identifying web community[C]//Proceedings of the Fourth International Workshop on Web Information and Data Management,United States,Nov 8 2002.United States:Association for Computing Machinery,2002:43-48.

[7] Asano Y,Nishizeki T,Toyoda M,et al.Mining communities on the web using a max-flow and site-oriented framework[J].IEICE Trans On Information and Systems,2006,E89-D:2606-2615.

[8] Ino H,Kudo M,Nakamura A.Partitioning of web graphs by community topology[C]//Proc of WWW2005,Japan.ACM,2005:661-669.

[9] Dourisboure Y,Geraci F,Pellegrini M.Extraction and classification of dense communities in the Web[C]//The 16th International World Wide Web Conference,Canada,May 8-12 2007.United States:Association for Computing Machinery,2007:461-470.

(上接 97 页)

的 ZigBee 路由算法,改进算法在传统 AODVjr 路由算法的基础上,结合 ZigBee 树路由算法,对路由发现过程中的 RREQ 分组进行控制,并且在数据传输过程中考虑节点剩余能量,并且及时对节点最小剩余能量值 E_{min} 进行调整,有效地节约了网络整体耗能,并使得网络的能量消耗达到均衡。仿真结果表明,该算法能有效地避开能量较低的节点进行数据的传输,节省了网络的总体能量消耗,延长了网络的寿命。但该算法没有考虑节点运动的情况,也没能提供对不同业务的 QoS 支持,这些方面正是未来的研究方向。

参考文献:

[1] ZigBee Document 053474r06[S].Version 1.0.Zi-gBee Alliance,2004.

[2] Baront P,Pillai P,Chook V W C.Wireless sensor networks:A survey on the state of the art and the 802.15.4 and ZigBee stand-stards[J].Computer Communications,2007,30(7):1655-1695.

[3] Chakeres I D,Klein-Berndt.AODVjr,AODV simplified [J].Mobile Computing and Communication Review,2002,6(3):100-101.

[4] Kim T,Kim D,Park N,et al.Shortcut tre-e routing in ZigBee network[EB/OL].[2008-02-16].http://resl.icu.ac.kr/~damiano/proc/iswpc2007_1.pdf.

[5] Ran Peng,Sen Mao-heng,Zou You-min.ZigBee routing selection strategy based ondata services and energy-balanced zigbee routing[C]//Proceedings of the 2006 IEEE Asia-Pacific Conference on Services Computing.Washington DC:IEEE Computer Society,2006:400-404.

[6] 耿萌.ZigBee 路由协议研究[D].郑州:解放军信息工程大学,2006.