

文章编号:1001-9081(2009)01-0213-04

一种改进的基于最大流的 Web 社区挖掘算法

张金增,范 明

(郑州大学 信息工程学院, 郑州 450052)

(mfan@zju.edu.cn)

摘要:针对原始最大流算法给每条边的边容量分配一个常量值,在社区质量及成员数量上造成的问题,提出了一种改进的 Web 社区挖掘算法。该算法考虑不同边的重要性差异,将加权 PageRank 算法中页面的重要度转化为衡量页面之间边重要性的传递概率值,并使用该值对边容量进行赋值。实验结果表明,改进的算法有效地提高了 Web 社区的质量。

关键词:Web 社区;Web 图;最大流算法;加权 PageRank

中图分类号: TP311.13 **文献标志码:**A

Mining Web community based on improved maximum flow algorithm

ZHANG Jin-zeng, FAN Ming

(College of Information Engineering, Zhengzhou University, Zhengzhou Henan 450052, China)

Abstract: Given that the original maximum flow algorithm set a fixed edge capacity to each edge, which caused poor quality and improper size of communities, this paper proposed an improved algorithm for mining Web communities. The algorithm considered the differences between edges in terms of importance, and assigned different capacities to different edges by transforming the significant measurements of pages evaluated by weighted PageRank algorithm to edge-transferring probability scores to measure the importance of edges, and assigning them to corresponding edges as their capacities. The experimental results show that the improved maximum flow algorithm improves the quality of Web community effectively.

Key words: Web community; Web graph; maximum flow algorithm; weighted PageRank

0 引言

随着 Internet 的快速发展,Web 资源飞速增长,并朝着多元化、复杂化的方向发展。如何有效地利用和发现 Web 中的有用信息成为难题。因此,从大量的信息中提取出与某一特定主题相关的 Web 页面变得异常重要。

Web 社区可以有效的发现与某一特定主题密切相关的 Web 页面集合。通常,将 Web 看作一个有向图,称之为 Web 图,其中图中每个顶点表示 Web 中的一个页面,图中的每条边表示页面之间的链接。一个 Web 社区是一个 Web 图的子图,发现社区的过程就是从该 Web 图中找到一个适当的割集。目前的社区发现技术^[1]大致有三种。文献[2]通过用 HITS 方法分析链接结构来获得 Web 社区,认为社区是由中心页面链接起来的,很稠密的权威页面构成的核。文献[3]从二分有向图的角度对社区给出了一种明确的定义描述,把 Web 社区看作一些二分有向图的核。文献[4-5]最先提出通过最大流算法发现 Web 社区,认为社区是具有社区内页面之间的链接数大于同社区外页面之间的链接数这一特性的页面形成的集合。

基于最大流的 Web 社区发现方法^[4-5]常常把包含噪声页面的图结构提取出来,并且在某些情况下不能提取出大小合适的社区。导致这些问题最主要原因是将边容量设置为常值。为了解决这些问题,本文提出了一种改进的最大流算法,使用加权 PageRank 算法^[6]中边的传递概率值对边容量进行动态赋值。在随机选择的 20 个主题上进行实验,实验结果表

明,本文提出的算法有效地降低了提取出噪声页面的可能性,提取出了更多与主题密切相关的有价值的页面,从而提高了所发现的 Web 社区的质量;此外,与原始算法相比,该算法所获得的社区中与主题相关的平均页面数目有明显提高。

1 相关工作及其存在的问题

1.1 最大流/最小割

网络 $s-t$ 最大流问题的定义如下:设 $G = \langle V, E \rangle$ 是有向图, $\langle u, v \rangle$ 是从节点 u 到 v 的有向边。设 $s, t \in V$ 是图 G 的源点(入度为 0)和汇点(出度为 0)。为每一条边 (u, v) 分配一个边容量 $c(u, v) \in \mathbf{Z}^+$, 该边上的流量 f 满足 $0 \leq f(u, v) \leq c(u, v)$, 并且对于所有的节点 $v \in V - \{s, t\}$ 有流入 v 的流量等于流出 v 的流量, 即 $\sum_{(u_i, v) \in E} f(u_i, v) = \sum_{(v, u_0) \in E} f(v, u_0)$ 。

网络 $s-t$ 最大流问题就是找出从源点 s 经由 G 流至汇点 t 的最大流量。

设 $T = V/S$, 其中 $S \subseteq V$, 给定 $s \in S, t \in T$, 那么边集 $\{(u, v) \in E \mid u \in S, v \in T\}$ 叫做 $s-t$ 的一个割集, 最小割集是所有割集中割边的容量和最小的一个。文献[7]的最大流 / 最小割定理已经证明了, 在一个网络图 G 中, $s-t$ 的最大流的流量等于分割 $s-t$ 的最小割集的容量。在所有的 $s-t$ 最大流算法中, 本文所使用的文献[8]提出的最短增广路径算法是一种简单有效的算法, 该方法非常适合用于发现 Web 社区, 其时间复杂度为 $O(VE^2)$ 。

1.2 最大流的 Web 社区发现算法

文献[4]根据图形理论, 从另一个角度提出了发现社区

收稿日期:2008-07-16;修回日期:2008-09-13。 基金项目:国家自然科学基金资助项目(60773048)。

作者简介:张金增(1983-),女,河南南阳人,硕士研究生,主要研究方向:数据挖掘、机器学习; 范明(1948-),男,河南信阳人,教授,博士生导师,CCF 会员,主要研究方向:数据库、数据挖掘、机器学习。

的方法。将社区定义为在 Web 图中具有这样一些特性的页面的集合,社区内的页面之间的链接(在两个方向)的密度要大于社区之间页面链接的密度。文献[5]证明了使用 $s-t$ 最大流算法提取的 Web 页面恰好满足 Web 社区内页面间的链接比社区外的页面链接要多的这一性质,并根据 $s-t$ 最大流算法设计了 Web 社区发现算法。

图 1 显示了从图(右边)中分离出的社区(左边)的例子。在图 1 中,如果 s 和 t 分别是左边和右边节点集中的节点,那么 $s-t$ 最大流算法返回从 s 到 t 的最大流和最小割(图 1 中有 5 条连接左边和右边节点集的边)。

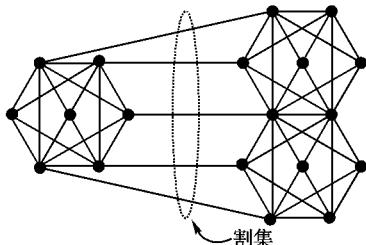


图 1 一个简单的 Web 社区例子

1.3 算法存在的问题

最大流社区发现算法虽然较好地解决了以往算法存在的主题漂移问题,但在获取社区过程中存在着一些问题。由于该算法给各边的边容量赋一个相同的常值,该值等于种子节点的个数。一方面造成了某些包含噪声页面的图结构常常被提取出来;另一方面两个与主题密切相关的页面之间的有价值链接变成了一条割边;此外,在一些情况下,每一次算法的迭代过程都不能获得新的成员页面,使得得到的 Web 社区的体积太小。

文献[9]对边容量和社区的体积之间的关系进行了深入的研究,得出可以通过增加边的容量来增大社区的体积。然而只通过增加边的容量并不能很好地解决这些问题,在很多情况下边容量增加的越大,主题漂移的问题就越容易显现出来,并且社区中噪声页面的数量也会不断增加。其主要的原因是没有考虑到每条边的不同重要程度,体现不出不同边的重要度对形成的社区的影响。为了解决这一问题,本文提出了一种有效地反映边重要性的边容量分配方法。

2 边容量分配思想

如上所述,为边容量分配一个常数值造成了社区在数量和质量上的一些问题。一种理想的解决办法就是,给重要的边分配一个更大的容量,不重要的边分配一个较小容量。那么关键的问题是如何定义一个重要边和不重要边,以及如何给不同重要程度的边分配不同的边容量,这就等于要给边定义一个等级或者分值,所以必须把它对页面的价值分布转化到对边的价值分布上来。

在基于链接分析的算法中,加权 PageRank 算法^{[6][30]}是用来计算页面分值最典型的方法,它建立在基于重要度传递关系构造的模型上,并对每个页面赋予一个衡量其重要性的权威值。该算法把页面的权威值通过链接页面之间的边转移概率值到与它相连的页面,并且为每个链接分配不同的权重,将更大的权重分配给与更重要的页面相连的超链接,这样与重要页面相连的边自然就会被分配更大的传递概率值。本文依据这个思想将页面的权威值转化为边的传递概率值,使用边的传递概率值去评估一条边的重要性,并用该值对边容量进行赋值。

加权 PageRank 算法页面权威值计算公式为:

$$PR(v) = \frac{d}{n} + (1-d) \times \sum_{u \in B(v)} PR(u) W_{(u,v)}^{in} W_{(u,v)}^{out} \quad (1)$$

其中:

$$W_{(u,v)}^{in} = \frac{I_v}{\sum_{p \in R(u)} I_p} \quad (2)$$

$$W_{(u,v)}^{out} = \frac{O_v}{\sum_{p \in R(u)} O_p} \quad (3)$$

$B(u)$ 是页面 u 的链入页面集, $R(u)$ 是页面 u 的链出页面集, I_u 是页面 u 的入链接数, O_u 是页面 u 的出链接数, d 是衰减因子, $0 < d < 1$, 通常取值为 0.15, n 为有向图 G 中节点的数量。

由式(1)可以推导出边的传递概率值公式为:

$$ER(u,v) = \frac{d}{n} + (1-d) \times PR(u) \times W_{(u,v)}^{in} \times W_{(u,v)}^{out} \quad (4)$$

在加权 PageRank 算法里得到的 ER 值在 $0 \leq ER(u,v) \leq 1$ 这个范围内,不适合作为边的容量,因为最大流算法里边容量必须是一个正整数,所以需要通过设定一个常量系数 f_q ,从而得到边容量的计算公式:

$$c(u,v) = f_q \times ER(u,v) \quad (5)$$

其中, f_q 是原始最大流社区发现算法中边容量的最大值,此时邻近图中的所有节点都在社区中, f_q 的值可以用与文献[9]中相同的方法求出。以上公式所得到的边容量是一个动态的变量。下面我们给出基于以上边容量分配方法的最大流社区发现算法的详细步骤。

3 改进的最大流社区发现算法

基于边的传递概率值的边容量分配方法的最大流社区发现算法的详细步骤如下:

- | | |
|--|-----------|
| 输入: $S = \{v_{s_1}, v_{s_2}, \dots, v_{s_l}\}$; | 种子节点集合 |
| 输出: $C = \{v_{c_1}, v_{c_2}, \dots, v_{c_m}\}$; | 一个 Web 社区 |
| 方法: | |
| 1) repeat | |
| 2) 围绕每一个种子节点 $v_{s_i} \in S$, 抽取深度为 2 的 Web 子图 $G(V, E)$; | |
| 3) 应用式(1)计算每个节点的 PageRank 值; | |
| 4) 增加源点 s 和汇点 t 到 V 中; | |
| 5) for 每个节点 $v_{s_i} \in S$ do | |
| 6) 增加边 (s, v_{s_i}) 到 E 中; | |
| 7) $c(s, v_{s_i}) = \infty$; | |
| 8) for 每条边 $(u, v) \in E$ do | |
| 9) 应用式(5)设置边容量 $c(u, v)$; | |
| 10) if $(v, u) \notin E$ then | |
| 11) 增加边 (v, u) 到 E 中; | |
| 12) $c(v, u) = c(u, v)$; | |
| 13) for 每个节点 $v \in V$ 且 $v \notin S \cup \{s, t\}$ do | |
| 14) 增加边 (v, t) 到 E ; | |
| 15) $c(v, t) = 1$; | |
| 16) 执行 $s-t$ 最大流算法; | |
| 17) 得到 $C = \{v \mid v \in V, \text{ 节点 } v \text{ 和种子节点连通}\}$; | |
| 18) 应用式(6)(将在后面介绍)计算每个节点 $v_{c_i} \in C$ 的分值; | |
| 19) 按分值大小对 C 中的节点进行排序; | |
| 20) 增加分值最高的一些非种子节点到 S 中; | |
| 21) until C 中的节点趋于稳定 | |

本文采取了与文献[4] 和 [5] 不同方法对 C 中的成员节

点计算分值。设 $v_{c_i}^{(in)}$ 表示从其他节点链接到它的入链接数, $v_{c_i}^{(out)}$ 表示它的出链接数。 $S_c(v_{c_i})$ 用来表示节点 v_{c_i} 的分值, 则计算分值的公式如下:

$$S_c(v_{c_i}) = PR(v_{c_i}) \times (v_{c_i}(\text{In}) + v_{c_i}(\text{Out})) \quad (6)$$

在文献[4]中计算分值仅仅以每个节点的链接数为依据, 因为有些排在前面分值较高的节点会有相同的链接数, 仅靠链接数来选择新的种子节点是不充分的, 本文把每个节点的 PageRank 值也考虑进去, 避免了这种情况。

4 实验结果及评价

4.1 实验数据的收集与清理

4.1.1 数据集

由于客观条件的限制, 不可能对所有网站数据进行研究。因此, 为了确保 Web 数据的获取不影响研究结果的可靠性, 选择了代表性的网页作为种子节点。在本文里, 使用一个简单的爬行程序收集数据集。查询等价于种子节点页面, 子图的构造从种子节点页面的爬取开始。节点的链出链接通过解析爬取到的网页的 HTML 文档得到, 链入连接通过 Google 搜索引擎得到。

4.1.2 种子节点集

本文选取了 20 个不同主题的网页作为种子节点页面, 并且每个种子节点的主题都具有明确的意义。此外, 在选取种子节点的时候, 如果种子节点有太多的链入或链出连接会导致边缘图过于庞大而引起社区中的主题漂移, 这样会降低所发现的 Web 社区的质量, 因此避免选取那些拥有太多链入或

链出链接的网页作为种子节点(入度小于 200)。

4.1.3 数据清理

数据清理是为了获得高质量的挖掘结果而做的准备工作。数据清理过程主要有以下几个方面:

1)首先排除入链接或者出链接数超过了 500 的 Web 页面, 因为这些页面往往是非常出名的一些站点页面, 像 Yahoo, Google 等, 这些站点页面根本就不需要使用挖掘策略就可以得到。

2)去除 URL 中包含%, ?, bbs, cgi-bin, diary, news 等的页面, 这些页面往往和用户要找的主题无关。

3)合并镜像页面, 所谓的镜像页面是指与主网站的内容相同的其他位置的网站页面就叫做镜像网站页面。

4.2 实验结果分析

为了验证本文提出算法的有效性, 选取了 20 个不同的主题进行实验, 对原始的算法和改进的算法进行了比较和分析。在下面的实验结果分析中分别用 C_1 表示原始最大流社区发现算法所得到的 Web 社区, C_2 表示改进算法得到的 Web 社区。

表 1 显示了对于每个种子节点, 所获得的社区 C_1, C_2 的大致情况。表 1 中前三列依次表示主题编号, 种子节点的 URL 和主题, $|V|$, $|C_1|$, $|C_2|$ 依次表示邻近图中节点的个数, 原始算法获得的社区成员数和改进算法获得的社区成员数。从表 1 中可以看出, 改进算法所获得的社区 C_1 在数量上好于原始算法, 编号为 2, 9, 20 这三个主题无论我们如何改变常值边容量, 都不能获得理想大小的 Web 社区。

表 1 种子节点、主题、邻近图节点个数和社区体积的相关情况

No.	Seed URLs	Topics	$ V $	$ C_1 $	$ C_2 $
1	http://www.sciam.com/	Science	2113	22	39
2	http://www.aaai.org	Artificial Intelligence	1384	9	29
3	http://www.ca.gov/	California	6175	36	47
4	http://www.epa.gov.cn/	Environmental Protection Agency	4772	64	89
5	http://succulent-plant.com/	Succulents	2390	21	29
6	http://www.ncac.org	Censorship	980	18	35
7	http://www.olympic.org/	Olympic	557	17	31
8	http://www.rockclimbing.com/	Rock climbing	2160	24	38
9	http://www.jaguarcars.com	jaguar	1390	4	27
10	http://www.gulfwarvets.com	Gulf war	603	17	26
11	http://www.jcrb.com/	法律法规	3939	66	84
12	http://junshi.xilu.com/	中国军事	2393	31	44
13	http://www.fec.com.cn/	财经证券	5553	28	47
14	http://www.yishu.com/	艺术/鉴赏	970	21	38
15	http://www.shufa.org/	书法	3261	58	64
16	http://www.5ijk.net/	医疗保健	3459	13	21
17	http://www.chinacars.com/	汽车	7247	14	85
18	http://www.edu.cn/	教育	4989	18	24
19	http://www.zhb.gov.cn/	环保	1627	19	32
20	http://www.lknet.ac.cn	园林	387	2	24

对改进前后两种算法所得到的社区 C_1 和 C_2 中前 15 个页面与主题相关的情况进行了比较, 实验结果如图 2 所示。从图中可以看出社区 C_1 和 C_2 的平均相关页面数分别为

5.95(范围分布从 2 到 10) 和 9.9(范围分布从 7 到 15), C_2 显著好于 C_1 ; 同时可以看出, 在所选的 20 个主题中有 17 个主题 C_2 比 C_1 好, 1 个主题的情况相同, 只有两个主题情况 C_2 比 C_1

差。而且大多数包括在 C2 中的相关页面也都包括在 C1 中了。

表 2 列举了关于主题 1 的社区 C1 和 C2 中前 15 个成员页面的 URL 的具体情况, 表中“+”代表该页面与种子节点页面主题相关, “-”代表该页面与种子节点页面主题不相关。主题 1 的种子页面是关于“科学”方面的内容,C2 中只有两个链接与主题“科学”不相关, 其余的 13 个页面均与主题密切相关。C1 中除了种子页面外和主题相关的页面数只有 4 个。C1 中排在第一的页面拥有到种子页面及其余页面的 34 个链接。当边容量分配为 9 的时候, 该页面及噪声页面均被提取出来, 其中 C1 中包含的 9 个不相关页面(3 ~ 5, 8 ~ 11, 13 ~ 14)都是从该页面链出的。本文给出的边容量分配方法降低了该页面与不相关

页面链接的边容量, 避免提取出这些噪声页面。从试验结果可以看出, 本文给出的改进算法所发现的社区成员的数量比原始的最大流算法所发现的社区成员多并且相关性更高。

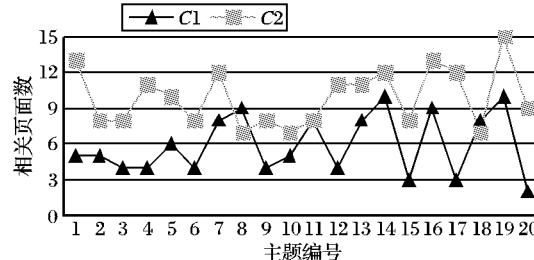


图 2 社区排名前 15 的网页中与主题相关的网页数量

表 2 关于主题“Science”原始算法和改进算法获得的社区 C1 和 C2 情况比较

社区 C1 的成员页面		社区 C2 的成员页面	
1 http://www.science.gov/	+	http://www.sciam.com/	+
2 http://www.usajobs.opm.gov/	-	http://www.science.gov/	+
3 http://www.sciam.com/	+	http://science.discovery.com/	+
4 http://www.grants.gov/	-	http://www.eere.energy.gov/	-
5 http://www.usa.gov/	-	http://worldwidescience.org	+
6 http://www.osti.gov/fedrnd/	+	http://www.osti.gov/	+
7 http://sciencediversitycenter.org/	+	http://www.sciencedaily.com/	+
8 http://www.whitehouse.gov/omb/egov/	-	http://www.usgs.gov/	+
9 http://www.cde.gov/ncidod/dhqar/ar_mrsa.html	-	http://www.nsf.gov/	+
10 http://www.regulations.gov/	-	http://www.quantcast.com/p-5b6qhwG1mz5To	-
11 http://www1.eere.energy.gov/education/	-	http://www.60secondscience.com/	+
12 http://www.scienceaccelerator.gov/	+	http://www.newsdaily.com/news/science/	+
13 http://www.ntweek.org	-	http://science-community.sciam.com/	+
14 http://www.govbenefits.gov	-	http://sciencediversitycenter.org/	+
15 http://www.orst.edu	-	http://www.osti.gov/scienceconferences/	+

5 结语

本文对原始的最大流社区发现算法进行探讨, 指出了为边容量赋常值原始算法中存在的问题。提出了一种改进的最大流发现社区算法, 依据边的不同重要性为边容量赋值, 即利用加权 PageRank 中边的传递概率值为边动态的分配容量。为了验证改进算法的有效性, 随机选择了 20 个种子页面进行了实验, 实验结果表明基于加权 PageRank 边传递概率的社区发现方法所提取的与主题相关的页面数平均是原始算法的 1.66 倍, 因此使用改进的算法发现 Web 社区是显著有效的。

下一步的工作将进一步改进边的容量, 并利用 Web 页面的内部结构将页面划分为基于单个主题的块, 在块的粒度上利用链接结构发现层次社区。

参考文献:

- [1] 高琰, 谷士文, 唐璇. 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, 23(7): 183 ~ 186.
- [2] GIBSON D, KLEINBERG J M, RAGHAVAN P. Inferring Web communities from link topology [C]//Proceedings of the 9th ACM Conference on Hypertext and Hypermedia. Pittsburgh: ACM Press, 1998: 225 ~ 234.
- [3] KUMAR R, RAGHAVAN P, RAJAGOPALAN S, et al. Trawling the Web for emerging cyber-communities [C]//Proceedings of the 8th International World Wide Web Conference. Toronto: Elsevier Science Press, 1999: 403 ~ 415.
- [4] FLAKE G W, LAWRENCE S, GILES C L, et al. Efficient identification of Web communities [C]// Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000: 150 ~ 160.
- [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] XING W P, GHORBANI A. Weighted PageRank Algorithm [C]// Proceedings of the 2nd Annual Conference on Communication Networks and Services Research. Fredericton, Canada: IEEE Computer Society, 2004: 305 ~ 314.
- [7] FORD L R, FULKERSON D R. Maximal flow through a network [J]. Canadian Journal of Mathematics, 1956 (8): 399 ~ 404.
- [8] EDMONDS J, KARP R. Theoretical improvements in algorithmic efficiency for network flow problems [J]. Journal of the ACM, 1972, 19(2): 248 ~ 264.
- [9] IMAFUJI N, KITSUREGAWA M. Finding a Web community by maximum flow algorithm with HITS score based capacity [C]// Proceedings of the Eighth International Conference on Database Systems for Advanced Applications, Washington, DC: IEEE Press, 2003: 101 ~ 106.