

电子超市网站链接结构优化的多目标模型

王有为¹, 汪定伟²

(1. 复旦大学管理学院, 上海 200433; 2. 东北大学信息科学与工程学院, 辽宁 沈阳 110004)

摘要: 链接结构是网站设计与维护时需要考虑的主要因素. 将网站结构抽象为无权有向图, 提出了网页可达性和平均载入时间的概念, 建立了一种网站链接结构优化的多目标模型, 并针对模型的非线性特点设计了禁忌搜索算法. 仿真实验表明, 链接结构优化的方法不但有助于管理员维护网站结构, 而且可以为电子超市的顾客获取商品信息带来方便.

关键词: 网页可达性; 平均载入时间; 电子超市网站; 链接结构优化; 多目标模型

中图分类号: TP391 **文献标识码:** A

Multi-objective model for link structure optimization of e-supermarket website

WANG You-wei¹, WANG Ding-wei²

(1. School of Management, Fudan University, Shanghai 200433, China;

2. School of Information Science and Engineering, Northeastern University, Liaoning Shenyang 110004, China)

Abstract: Link structure is one of the key factors that must be considered when designing and authoring Web sites. By considering the site link structure as a weightless digraph, the concepts of web page accessibility and average loading time were presented. Then, a multi-objective model was built and a tabu search algorithm was proposed to solve this NLP. Simulation results revealed that the method introduced could help the e-supermarket customers to access information more comfortably and assist the site authoring at the same time.

Key words: web page accessibility; average loading time; e-supermarket website; link structure optimization; multi-objective model

1 引言(Introduction)

电子超市是 B-to-C 电子商务的一种表现形式, 经营电子超市的公司主要通过网站进行商品的宣传和交易. 为满足不断变化的需求, 网站需要跟踪顾客的行为并适时调整网站的链接结构, 才能在竞争中保持有利地位. 网站结构设计问题已经引起众多学者的关注^[1-5]. Nakayama 等^[1]提出一种发现顾客的访问行为与网站设计者目标之间差异的方法, 可用于指导网站的优化设计. Garofalakis 等^[2]以网页的相对访问次数作为标准, 提出一种网站结构动态更新的方法. 但作者也指出, 此方法在网页之间存在较强链接约束时难以应用(如电子商务网站). Yen 和 Fu^[4]的方法通过分析网页的可达性, 可以指出网站设计的不合理之处. Kim 和 Yoo^[5]将网站上的链接分为 5 类, 并采用实验的方法研究了链接的组合应用

效果. 以上研究者对网站的结构设计提出了指导性的方法, 但在网站设计实践中, 管理员更需要一些具体的操作建议. 本文建立一种网站链接结构设计的多目标模型, 网站管理员通过考虑求解此模型的优化结果, 可以有效更新和优化网站的链接结构.

2 网站结构的图描述(Description of website structure by graph)

电子超市网站中的链接主要分为两类, 一类反映商品目录结构的链接称为基本链接, 其余为方便顾客浏览的链接为附加链接^[5]. 如果将网页和链接分别视为顶点和弧, 则网站结构可以抽象为一个带标号的有向图. 设网站中共有 N 个网页, 它们的标号为 0 到 $N-1$, 其中主页的标号为 0. 定义布尔矩阵 $B = \{b_{i,j}; i, j = 0, 1, \dots, N-1\}$. 其中, 若 $b_{i,j} = 1$ 代表链接 (i, j) 为基本链接, $b_{i,j} = 0$ 代表链接 (i, j) 为

附加链接. 定义布尔矩阵 $X = \{x_{i,j} \mid i, j = 0, 1, \dots, N-1\}$ 代表网站的一种链接结构. 其中若 $x_{i,j} = 1$ 代表链接 (i, j) 存在, $x_{i,j} = 0$ 代表链接 (i, j) 不存在.

链接不存在长短的差异, 因此网站结构图是一个无权图. 网页的层次定义为从主页到达网页所经过的最少链接个数^[4], 对于无权图而言可以用广度优先算法来计算^[6,7]. 图1是一个网站链接和网页层次的例子, 增、减链接都可能影响网页的层次.

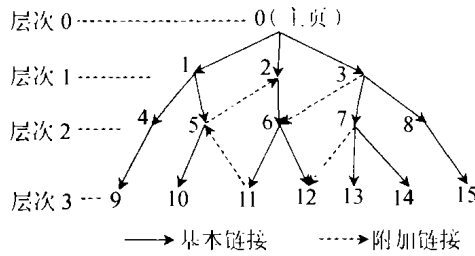


图1 网页的层次

Fig. 1 Levels of web page

3 基本概念(Basic conception)

3.1 链接的可达性(Accessibility of hyperlinks)

链接的可达性取决于其所在的网页的情况. 本文将链接的可达性定义为顾客点击此链接的可能性. 在不对任何链接加亮显示并且不考虑访问者不同兴趣、爱好的情况下, 每个链接被点击的可能性是相同的, 此时链接 (i, j) 的可达性可用下式计算:

$$AL_{i,j} = x_{i,j} / \sum_{j=0}^{N-1} x_{i,j}, \quad i, j = 0, 1, \dots, N-1. \quad (1)$$

3.2 网页的可达性(Accessibility of web pages)

网页的可达性是定义在网站层次上的^[4], 本文中将网页的可达性定义为用户沿所有路径到达此网页的可能性之和. 由于实际网站上的网页和链接数目很多, 计算所有可能的路径几乎是不可能的. 本文考虑最主要的路径, 即用户按照网页层次由浅到深的顺序访问的路径, 这样的路径一定包含了由主页到达此页面的最短路径. 为得到上述方式定义的到达每个网页的所有路径, 可以采用下面的路径树生成算法(path tree spanning algorithm, 简称 PTSA), 访问过程记录到一棵路径树中. 此算法流程如下:

1) 访问主页, 然后访问从主页可以直接访问的任一页面 W1;

2) 访问满足条件“层次比当前页面低而且可以由当前页面直接访问”的页面 W2;

3) 从页面 W2 开始重复以上相同的访问, 直到遇到一个没有满足上述条件的网页为止;

4) 沿以上相同的访问次序返回到一个仍存在

满足上述条件的网页, 再重复与步骤 3) 相似的访问直到路径树中所有网页都不满足上述条件为止.

按 PTSA 算法, 图2中到页面1的可达路径为 $\{0, 1\}$; 到页面2的可达路径为 $\{0, 2\}$, 到页面3的可达路径为 $\{0, 1, 3\}$, $\{0, 2, 3\}$, 到页面4的可达路径为 $\{0, 1, 4\}$, $\{0, 2, 4\}$.

一旦网站结构确定, 根据路径树下列参数可以计算:

$$N_i = f_i(X), \quad i = 1, 2, \dots, N-1, \quad (2)$$

$$L_{i,l} = \Phi_{i,l}(X), \quad i = 1, 2, \dots, N-1, \quad l = 1, 2, \dots, N_i, \quad (3)$$

$$\begin{cases} J_{i,l,j} = \Omega_{i,l,j}(X), \\ i = 1, 2, \dots, N-1, \quad l = 1, 2, \dots, N_i, \quad j = 1, 2, \dots, (L_{i,l}+1). \end{cases} \quad (4)$$

其中: N_i 为用户可以到达网页 i 的路径条数; $L_{i,l}$ 为到达网页 i 的第 l 条路径所需的步数; $J_{i,l,j}$ 为到达网页 i 的第 l 条路径的第 j 个网页的标号. 于是网页 i 的可达性可由下式计算:

$$AP_i = \sum_{l=1}^{N_i} \prod_{j=1}^{L_{i,l}+1} AL_{J_{i,l,j}, J_{i,l,j+1}}, \quad i = 1, 2, \dots, N-1. \quad (5)$$

本文假设顾客都是首先访问网站主页(即此时已达到主页), 因此不定义主页的可达性.

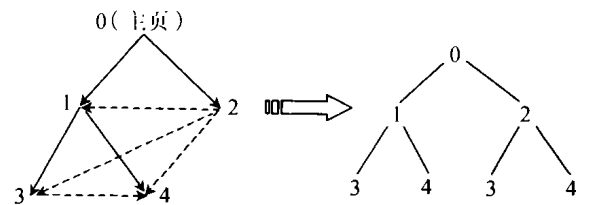


图2 应用路径树生成算法的例子

Fig. 2 Example of applying PTSA

3.3 平均载入时间(Average loading time)

设 $T_i (i = 0, 1, \dots, N-1)$ 为网页 i 的平均下载时间, 它与网页大小以及平均网络传输速度有关. 一个网页的平均载入时间 TL_i 与到达此网页的路径条数和路径上网页的载入时间的关系如式(6)所示.

$$TL_i = \frac{1}{N_i} \left(\sum_{l=1}^{N_i} \sum_{j=1}^{L_{i,l}+1} T_{J_{i,l,j}} \right), \quad i = 1, 2, \dots, N-1. \quad (6)$$

3.4 网页访问率(Visiting rate of web pages)

根据网络服务器在过去一段时间内统计得到的每一网页被访问的次数, 可以按下式计算网页访问率(与上节同样的原因, 不定义主页的访问率):

$$Q_i = V_i / \sum_{i=1}^{N-1} V_i, i = 1, 2, \dots, N-1. \quad (7)$$

其中: Q_i 为网页 i 的用户访问率, V_i 为过去一段时间内用户访问网页 i 的次数.

4 链接结构优化模型(Model of link structure optimization)

4.1 模型的建立(Establishment of mathematical model)

按照方便顾客的原则,网站在进行链接调整时应使访问率高的网页保持较大的可达性和较少的载入时间.本文中用相关性来衡量两组量 $U_i, V_i (i = 1, 2, \dots, N-1)$ 保持一致的程度(关于相关性详见文献[1]).

$$\text{cov}(U, V) = \frac{\sum_{i=1}^{N-1} (U_i - U)(V_i - V)}{\sqrt{\sum_{i=1}^{N-1} (U_i - U)^2 \cdot \sum_{i=1}^{N-1} (V_i - V)^2}}.$$

为保证网站整体结构的稳定性,在一次调整时变化不能太大.所以在每个页面上增加或减少的链接个数不能太多,同样也应防止新增加的链接过多的集中于少数网页或者减少链接时将指向某个网页的链接过多地删除.另外,在增加链接时应保证被连接的网页在内容上存在相关性,否则容易造成网站的业务逻辑混乱.用布尔矩阵 $U = \{u_{i,j} | i, j = 0, 1, \dots, N-1\}$ 代表网页间是否存在相关性,其中若 $u_{i,j} = 1$ 代表网页 i 和网页 j 内容相关,否则不相关.显然 U 为对称矩阵.用布尔矩阵 $A = \{a_{i,j} | i, j = 0, 1, \dots, N-1\}$ 代表网站初始的链接结构.其中,若 $a_{i,j} = 1$ 代表链接 (i, j) 存在,若 $a_{i,j} = 0$ 代表链接 (i, j) 不存在.定义变量 $X = \{x_{i,j}\}$:

$$x_{i,j} = \begin{cases} 1, & \text{链接}(i, j) \text{ 存在,} \\ 0, & \text{其他,} \end{cases} \quad i, j = 0, 1, \dots, N-1. \quad (8)$$

于是建立网站链接结构优化模型如下:

$$\max f_1(X) = \text{cov}(AP(X), Q), \quad (9)$$

$$\max f_2(X) = \text{cov}(Tl(X)^{-1}, Q), \quad (10)$$

$$\text{s.t.} \sum_{j=0}^{N-1} |x_{i,j} - a_{i,j}| \leq R_i, \quad i = 0, 1, \dots, N-1, \quad (11)$$

$$\sum_{j=0}^{N-1} |x_{j,i} - a_{j,i}| \leq C_i, \quad i = 0, 1, \dots, N-1, \quad (12)$$

$$x_{i,j} - b_{i,j} \geq 0, \quad i, j = 0, 1, \dots, N-1, \quad (13)$$

$$u_{i,j} - x_{i,j} \geq 0, \quad i, j = 0, 1, \dots, N-1. \quad (14)$$

式(9)表示极大化网页可达性与访问率间的相关性;式(10)表示极大化网页载入时间的倒数与访问率间的相关性;式(11)为在每个网页上增加或减少链接的个数约束;式(12)为增加或减少指向每个网页的链接个数约束;式(13)表示基本链接不可以被删除;式(14)表示新增加链接所涉及的两个网页的内容必须存在相关性.

根据式(1)~(7),模型(8)~(13)中的 $AP_i, Tl_i (i = 1, 2, \dots, N-1)$ 为变量 X 的函数, $Q_i (i = 1, 2, \dots, N-1), R_i, C_i (i = 0, 1, \dots, N-1)$ 以及 $a_{i,j}, b_{i,j}, u_{i,j} (i, j = 0, 1, \dots, N-1)$ 为常量.

4.2 转化为单目标问题(Conversion to single objective problem)

设多目标问题(8)~(14)的理想点为 (f_1^*, f_2^*) , 并令 $E(X)$ 为式(11)~(14)构成的可行域.采用处理多目标问题的极大模理想点法^[8],上述问题可以转化为求解如下单目标问题.

$$\min \lambda \quad (15)$$

$$\text{s.t. } X \in E(X), \quad (16)$$

$$f_1^* - f_1(X) \leq \lambda, \quad (17)$$

$$f_2^* - f_2(X) \leq \lambda, \quad (18)$$

$$\lambda \geq 0. \quad (19)$$

根据极大模理想点法的相关定理,问题(15)~(19)的最优解为原问题(8)~(14)的弱有效解.

5 模型求解(Solving methods of model)

5.1 禁忌搜索算法(Tabu search)

在计算可达路径时要用到 PTSA 算法,使得网页可达性和载入时间无法用解析形式表达,故上述模型属于非线性规划,难以用常规方法处理.本文设计 TS(关于 TS 详见文献[9,10])求理想点以及问题(15)~(19),其关键环节设计如下:

1) 初始解.初始解与网站现有结构相同,即 $x_{i,j} = a_{i,j} (i, j = 0, 1, \dots, N-1)$. 按照模型(8)~(14),此解一定满足约束条件.

2) 领域结构.当前解的领域定义为任意改变变量矩阵中的一位 (i, j) (一次移动)所能达到的解的集合.领域中的解还要满足约束条件,不满足约束条件的解令其目标函数值为 0.

3) 禁忌表与长期表.禁忌表记录刚刚经过的 TabuSize 次移动,本文中 TabuSize = 7.在长期表 F_{max} 中记录每个移动 (i, j) 发生的次数,并施加频数惩罚: $f'(X) = f(X) - w \cdot \text{penalty}(i, j)$.

4) 吸收水平函数.当禁忌表中某个解的目标函

数值超过历史上的最好解时,取消禁忌.

5) 停止准则. 当迭代次数超过 N_MAX 时, 算法停止.

5.2 计算举例 (Case study)

图3为一个网站的基本链接结构,附加链接见表1,在过去一段时间内每个网页的访问次数如表2,平均下载时间见表3.表4中给出了矩阵 U 中为0的元素集合,即不相关的网页对.

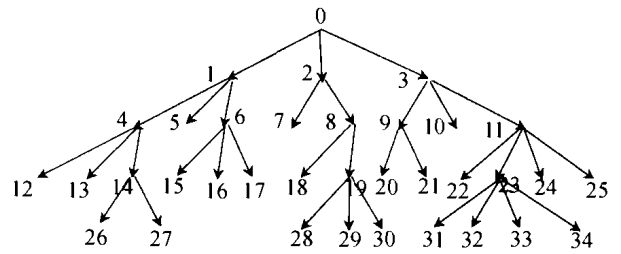


图 3 网站的基本链接结构

Fig. 3 Basic link structure of website

表 1 优化前网站的附加链接

Table 1 Add-on links before optimization

(0,4)	(0,18)	(0,23)	(0,26)	(2,6)	(2,17)	(8,16)	(9,14)	(15,27)	(17,30)
(20,31)	(21,34)	(23,30)	(25,26)	(26,0)	(26,3)	(27,1)	(27,3)	(28,1)	(28,4)
(29,3)	(30,0)	(30,5)	(31,5)	(31,7)	(32,5)	(32,7)	(32,8)	(33,9)	(34,6)

表 2 网页的访问次数

Table 2 Visiting times of pages

V_i	0	1	2	3	4	5	6	7	8	9
- -	120	19	38	36	72	52	61	33	15	23
10 +	23	13	55	63	34	43	15	16	7	12
20 +	21	38	14	64	24	35	92	51	11	22
30 +	62	41	32	22	36					

表3 网页的平均下载时间

Table 3 Average downloading time of pages

T_i/s	0	1	2	3	4	5	6	7	8	9
- -	5.28	3.64	0.26	6.20	1.50	9.94	2.76	5.44	0.88	8.56
10 +	1.24	16.04	4.14	0.78	0.74	0.78	7.26	0.88	27.64	0.54
20 +	1.44	0.90	2.62	3.88	2.54	2.50	2.92	2.96	1.56	1.66
30 +	2.18	2.90	1.72	3.20	2.14					

表 4 内容不相关的网页对

Table 4 Irrelative page couples

源网页	目标网页	源网页	目标网页
1	2,3,7,18,20,22,24,25,29,30,33	16	17,28,30
2	3,5,6,9,11,14,15,21,26,27,31,32	17	18,19,21,24,26,29,34
3	4,6,7,13,17,18,19,28,30	18	20,23,28,31
4	5,7,9,13,16,18,20,23,30,31,32,34	19	23,30,32
5	7,10,11,17,23,29,32,33	20	22,23,24,26,29,31
6	7,8,12,18,21,30,32	21	24,29,32
7	9,11,15,17,21,29,30,33	22	26,33,34
8	11,13,21,22,23,25,27	23	27
9	17,19,23,26,29,30,31,33,34	24	25,26,29
10	15,20,22,28,29,31,32	25	29,30,33
11	14,19,29,30	26	29,33,34
12	14,17,19,28	27	30,32,34
13	16,27	28	29,30,31,32
14	15,32	29	33,34
15	26,29,32	30	31,33,34

本文用 Java 语言实现了以上算法,在 $R_i = C_i = 2 (i = 0, 1, \dots, N - 1)$ 时进行了仿真实验,得到的仿真结果见表 5。结构调整后目标函数值 f_1 由 0.381 增加到 0.736, f_2 由 0.251 增加到 0.672,改进效果明显。

表 5 理想点和弱有效解

Table 5 Ideal points and weak efficient solutions

	初始解	理想点	弱有效解	与理想点之差
$f_1(X)$	0.381	0.796	0.736	0.061
$f_2(X)$	0.251	0.757	0.672	0.085*

“*”:问题(15)~(19)的最优解.

图4和图5分别是网页可达性和载入时间的倒数相对于访问率的变化情况。

优化后附加链接的变化情况见表6,其中“+”和“-”分别表示增加和减少的链接。

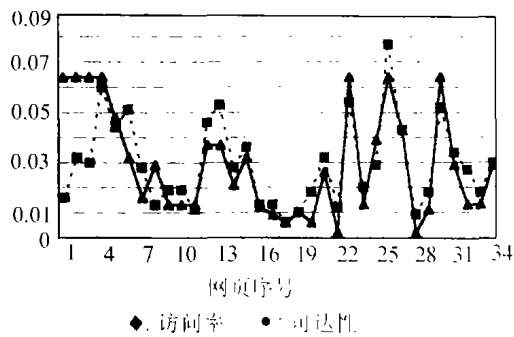


图4 网页可达性与访问率的变化情况 (cov = 0.736)

Fig. 4 Relationship of page accessibility and visiting rate (cov = 0.736)

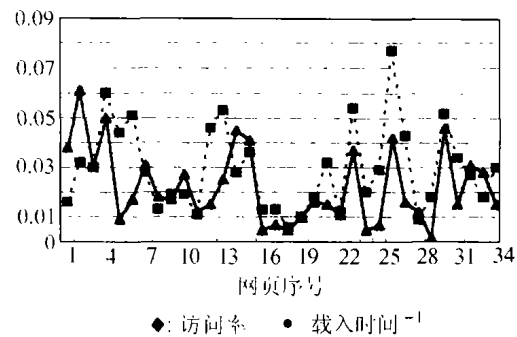


图5 网页载入时间与访问率的变化情况 (cov = 0.672)

Fig. 5 Relationship of page loading time and visiting rate (cov = 0.672)

表6 优化后增加和删除的附加链接

Table 6 Added and deleted add-on links after optimization

+ (0,30)	+ (1,34)	+ (2,13)	+ (3,8)	+ (3,21)	+ (5,8)	+ (6,1)	+ (6,24)	+ (7,12)	+ (7,32)
+ (8,0)	+ (8,1)	+ (10,9)	+ (10,33)	+ (11,17)	+ (11,28)	+ (12,2)	+ (12,3)	+ (13,25)	+ (16,5)
+ (16,6)	+ (17,4)	+ (17,6)	+ (18,29)	+ (20,5)	+ (20,7)	+ (21,4)	+ (21,9)	+ (22,29)	+ (23,21)
+ (24,7)	+ (24,10)	+ (25,11)	+ (25,13)	+ (26,12)	+ (26,31)	+ (27,11)	+ (27,14)	+ (28,15)	+ (28,19)
+ (29,16)	+ (29,18)	+ (30,15)	+ (31,23)	+ (31,26)	+ (32,23)	+ (32,26)	+ (33,19)	+ (33,24)	+ (34,31)
+ (34,32)	- (0,18)	- (2,17)	- (9,14)	- (23,30)	- (30,0)				

在以上实验中,算法的平均运行时间为 88.4 s,这在中小规模的网站结构优化问题中是可以接受的.考虑到 Java 算法的效率较低,以及网站结构可以采用邻接表等存储结构,因此在实践中算法的效率仍有提高的余地,这值得将来作进一步的研究。

6 结论 (Conclusion)

依据访问率并综合考虑网页的可达性与平均载入时间的多目标模型可以有效地改进网站的链接结构,本文提出的方法给设计者建设更加方便顾客访问的电子超市网站提供了一条可行的途径。

参考文献 (References):

- [1] NAKAYAMA T, KATO H, YAMANE Y. Discovering the gap between web site designers' expectations and users' behavior [J]. *Computer Networks*, 2000, 33(1-6): 823-835.
- [2] GAROFALAKIS J, KAPPOS P, MOURLOUKOS M. Web site optimization using page popularity [J]. *IEEE Internet Computing*, 1999, 3(4): 22-29.
- [3] WANG Y W, WANG D W. Design strategy of web page for e-supermarket [C] // JIANG Pingyu, HUANG G, CHENG Kai. *Proc of 2001 Int Conf on eCommerce Engineering*. Xi'an: China Machine Press, 2001.
- [4] YEN B P, FU K. Accessibility on web navigation [C] // JIANG Pingyu, HUANG G, CHENG Kai. *Proc of 2001 Int Conf on eCom-*

merce Engineering. Xi'an: China Machine Press, 2001.

- [5] KIM J, YOO B. Toward the optimal link structure of the cyber shopping mall [J]. *Int J of Human-Computer Studies*, 2000, 52(3): 531-551.
- [6] SHAFFER C A, 著. 张明, 刘小丹, 译. 数据结构与算法分析 [M]. 北京: 电子工业出版社, 2001.
(SHAFFER C A. Translated by ZHANG Ming, LIU Xiaodan. *Data Structure and Algorithm Analysis* [M]. Beijing: Electronics Industry Press, 2001.)
- [7] 潘道才. 数据结构 [M]. 成都: 成都电讯工程学院出版社, 1988.
(PAN Daocai. *Data Structure* [M]. Chengdu: Chengdu Telecommunication Engineering Press, 1988.)
- [8] 胡锐达. 实用多目标最优化 [M]. 上海: 上海科技出版社, 1990.
(HU Liuda. *Practical Multi-objective Optimization* [M]. Shanghai: Shanghai Science and Technology Press, 1990.)
- [9] GLOVER F. Tabu Search - Part I [J]. *ORSA Journal on Computing*, 1989, 1(3): 190-206.
- [10] GLOVER F. Tabu Search - Part II [J]. *ORSA J of Computing*, 1990, 2(1): 4-32.

作者简介:

王有为 (1974—), 男, 获东北大学系统工程专业博士学位, 现工作于上海复旦大学管理学院. 主要研究方向为电子商务中的建模与优化, 智能优化方法等. E-mail: ywwang@fudan.edu.cn;

汪定伟 (1948—), 男, 东北大学教授, 博士生导师. 主要研究方向为制造系统建模与优化, MRP-II/ERP, 智能计算与软计算方法等. E-mail: dwwang@mail.neu.edu.cn.