文章编号：1002-0411(2003)01-019-04

# 基于网页可达性和访问率的电子超市网站链接结构优化方法

王有为　汪定伟

（东北大学信息科学与工程学院　沈阳　110004）

**摘　要**：定义了链接可达性和网页可达性的概念. 为计算网页可达性，设计了计算到达网页路径的路径树生成算法(PTSA). 建立了一种极大化网页访问率与可达性之间相关性的网站链接结构调整的数学模型，并提出将 PTSA 嵌入禁忌搜索的求解方法. 试验结果表明本文的方法可以帮助网站设计者改进网站的链接结构.

**关键词**：网页可达性；网页访问率；电子超市网站；链接结构优化；禁忌搜索

**中图分类号**：TP391　　**文献标识码**：B

# ACCESSIBILITY AND VISITING RATE BASED OPTIMAL ADJUSTMENT APPROACH TO LINK STRUCTURE FOR E-SUPERMARKET WEBSITE

WANG You-wei　WANG Ding-wei

(*School of Information Science and Engineering, Northeastern University, Shenyang*　110004, *China*)

**Abstract**: Link accessibility and page accessibility are defined in this paper. To compute the page accessibility, a Path Tree Spanning Algorithm (PTSA) was introduced. A mathematical model was presented to maximize the covariance of visiting rate and accessibility of Web pages. As solving method, PTSA was embedded in the Tabu Search for optimal solution. Case study proved that the method presented in this paper could help the site designer to improve the link structure of E-Supermarket Websites.

**Keywords**: web page accessibility, web page visiting rate, e-supermarket Website, optimization of link structure, tabu search

## 1　引言(Introduction)

电子超市是 B-to-C 电子商务的一种重要表现形式，网站是其商品宣传和交易的主要窗口. 网站结构设计问题已经引起众多学者的关注[1,2,3,4,5]. 由于顾客对商品的需求和对网站的访问方式经常在变化，电子超市网站需要根据这种变化对其链接结构定期调整.

## 2　网站结构的图描述(Description of Site Structure by Graph)

电子超市网站中反映商品目录结构的链接可以称为基本链接，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，并为每个网页编号，则网站结构可以抽象为带标号的有向图. 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 $B=\{b_{i,j}\ i,j=0,1,\cdots,N-1\}$，其中若 $b_{i,j}=1$ 代表链接$(i,j)$为基本链接，$h_{i,j}=0$ 代表链接$(i,j)$为附加链接. 定义布尔矩阵 $X=\{x_{i,j}|i,j=0,1,\cdots,N-1\}$代表网站的链接结构，其中若 $x_{i,j}=1$ 代表链接$(i,j)$存在，$x_{i,j}=0$ 若代表链接$(i,j)$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，增加减少链接都可能使网页的层次发生

收稿日期：2002－02－25
基金项目：国家自然科学基金资助项目(60084003,70171056)

变化.

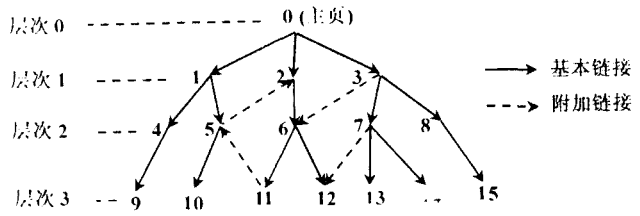


图 1 网页的层次

Fig. 1 Levels of Web page

## 3 基本概念(Fundamental Conceptions)

### 3.1 链接的可达性

链接的可达性是定义在页面层次上的[4]. 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$AL_{i,j}$可用式(1)计算.

$$AL_{i,j}=x_{i,j}/\sum_{j=0}^{N-1}x_{i,j} \qquad i,j=0,1,\cdots,N-1 \qquad (1)$$

### 3.2 网页的可达性

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. 但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. 本文只考虑最主要的路径,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. 为得到按上述方式定义的到达每个网页的所有路径,可采用下面的路径树生成算法(Path Tree Spanning Algorithm: PTSA),其流程如下:

1) 访问主页,然后访问从主页可以直接访问的任一页面 $P1$.

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 $P2$;

3) 从页面 $P2$ 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}.

定义下列符号:

$$N_i=f_i(X) \qquad i=1,2,\cdots,N-1 \qquad (2)$$

$$L_{i,l}=\Phi_{i,l}(X) \qquad i=1,2,\cdots,N-1;\ l=1,2,\cdots,N_i \qquad (3)$$

$$J_{i,l,j}=\Omega_{i,l,j}(X) \qquad i=1,2,\cdots,N-1;\ l=1,2,\cdots,N_i;j=1,2,\cdots,(L_{i,l}+1) \qquad (4)$$

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 $i$ 的可达性可用(5)式计算:

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

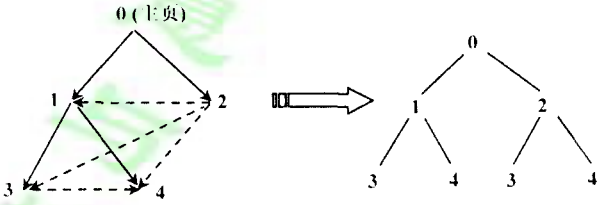


图 2 应用路径树生成算法的例子

Fig. 2 An example of applying PTSA

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.

### 3.3 网页访问率

网络服务器可以统计过去一段时间内每一网页被访问的次数,于是访问率可以按下式计算(与上节同样的原因,不定义主页的访问率):

$$Q_i=V_i/\sum_{i=1}^{N-1}V_i \qquad i=1,2,\cdots,N-1 \qquad (6)$$

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 数学模型(Mathematical Model)

网站在进行链接调整时应考虑使网页的可达性与其访问率保持一致,本文中用相关性来衡量[1]. 此外为保证网站整体结构的稳定性,一次调整时变化不能太大. 所以在每个页面上增加/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. 用布尔矩阵 $A=\{a_{i,j}|i,j=0,1,\cdots,N-1\}$代表网站初始的链接结构,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:

$$x_{i,j}=\begin{cases}1; & \text{链接}(i,j)\text{ 存在}\\ 0; & \text{其它}\end{cases} \qquad i,j=0,\cdots,N-1 \qquad (7)$$

于是网站链接结构调整问题可用模型(8－11)来描述：

$$\max f(x)=\left(\sum_{i=1}^{N-1}(AP_i-\overline{AP})(Q_i-\overline{Q})\right)\Big/\sqrt{\sum_{i=1}^{N-1}(AP_i-\overline{AP})^2\cdot\sum_{i=1}^{N-1}(Q_i-\overline{Q})^2} \quad (8)$$

Subject To：

$$\sum_{j=0}^{N-1}|x_{i,j}-a_{i,j}|\leqslant R_i \quad i=0,1,\cdots,N-1 \quad (9)$$

$$\sum_{j=0}^{N-1}|x_{j,i}-a_{j,i}|\leqslant C_i \quad i=0,1,\cdots,N-1 \quad (10)$$

$$x_{i,j}-b_{i,j}\geqslant 0 \quad i,j=0,1,\cdots,N-1 \quad (11)$$

式(8)表示极大化网页可达性和访问率间的相关性，式(9)为在每个网页上增加或减少链接的个数约束，式(10)为增加或减少指向每个网页的链接个数约束. 式(11)表示基本链接不可以被删除. 根据式(1－6)，模型(8－11)中的 $AP_i(i=1,2,\cdots,N-1)$ 为变量 $x_{i,j}$ 的函数，$Q_i(i=1,2,\cdots,N-1)$，$R_i$，$C_i(i=0,1,\cdots,N-1)$ 为常量.

## 5　禁忌搜索(Tabu Search)

在计算可达路径时要用到 PTSA 算法，使得网页可达性无法用解析形式表达，故上述模型难以用常规方法处理，本文用 TS[8,9] 来求解此问题，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 $x_{i,j}=a_{i,j}(i,j=0,1,\cdots,N-1)$.

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 $F|_{N\times N}$ 中记录每个移动$(i,j)$发生的次数，并施加频数惩罚.

4) 吸收水平函数：当禁忌表中某个解优于历史上的最好解时，取消禁忌.

5) 停止准则：当迭代次数超过 100 时，算法停止.

## 6　计算举例(Case Study)

### 6.1　实验数据

图 3 为一个网站的基本链接结构图，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. 优化前网页的访问率和可达性如图 4 所示，它们间的相关性为 0.381.

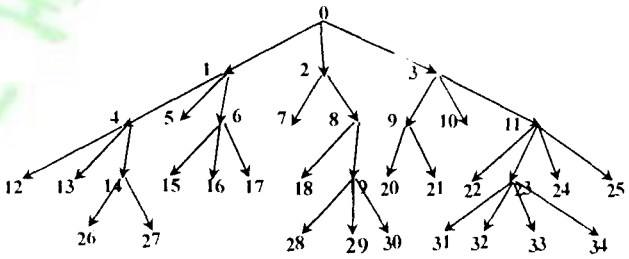


图 3　网站的基本链接结构

Fig. 3　Basic link structure of Website

表 1　优化前网站的附加链接

Tab. 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

Tab. 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 | | | | | |

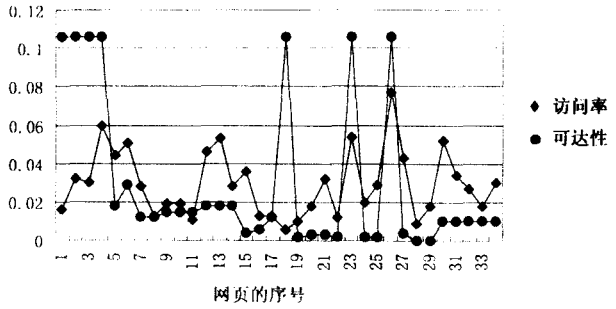


图 4 优化前网页的可达性及访问率(Cov=0.381)

Fig. 4 Accessibility and visiting rate of pages before optimization (Cov=0.381)

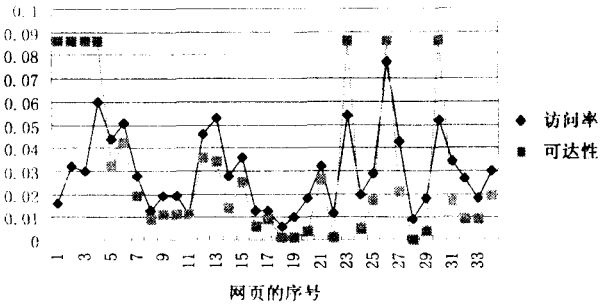


图 5 最多增减 2 个链接时优化后的网页可达性(Cov=0.801)

Fig. 5. Accessibility of pages when at most two links can be modified after optimization(Cov=0.801)

### 6.2 优化结果

图 5 是每个网页上最多增减两个链接时的优化结果,目标函数值为 0.801.

由图 5 可见,结构调整后网页访问率和可达性之间的相关性明显增大,达到了预期的优化目标.而且网站允许调整的范围越大,相关性的增加越显著.当每个网页上最多增减两个链接时的优化结果见表 3.

表 3 优化后增加和删除的附加链接

Tab. 3 Added and deleted add-on links after optimization

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

## 7 结论(Conclusions)

网页可达性与访问率间相关性的大小可以作为评价电子超市网站的链接结构是否合理的一个标准,本文提出的链接结构改进方法可以帮助网站设计者建设浏览更加方便的电子超市网站.

## 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**: 823～835

2 Garofalakis J, Kappos P, Mourloukos M. Web site optimization using page popularity [J]. IEEE Internet Computing, 1999, Jul. −Aug: 22～29

3 Wang Y W, Wang D W, Design strategy of web page for e−supermarket [A], Jiang Pingyu et. al, International Conference on eCommerce Engineering, Xi'an: China Machine Press, 2001, 101～107

4 Yen B P, Fu K. Accessibility on web navigation [A], Jiang Pingyu et. al, 2001 International Conference on eCommerce Engineering 2001, Xi'an: China Machine Press, 2001, 30～37

5 Kim J, Yoo B. Toward the optimal link structure of the cyber shopping mall [J]. Int. J. Human-Computer Studies, 2000, **52**: 531～551

6 C. A. Shaffer 著,张铭 刘晓丹译. 数据结构与算法分析(Java 版)[M]. 北京:电子工业出版社,2001

7 潘道才. 数据结构[M]. 成都:成都电讯工程学院出版社,1988

8 F. Glover, Tabu Search-Part I, ORSA Journal on Computing 1 (1989) 190～206

9 F. Glover, Tabu Search-Part II, ORSA Journal on Computing 2 (1990) 4～32


## 作者简介


**王有为**(1974−),男,东北大学信息学院博士研究生.研究领域为电子商务中建模与优化,智能优化方法等.

**汪定伟**(1948−),男,教授,博士生导师.研究领域为制造系统建模与优化,MRP−II/ERP,智能计算与软计算方法等.