

一种基于页面聚类的 Web 概念化建模新方法

姚 瑶¹, 王战红², 石 磊³

(1 中州大学 信息工程学院, 河南 郑州 450044; 2 铁道警察学院 公安技术系, 河南 郑州 450053;

3 郑州大学 信息工程学院, 河南 郑州 450001)

摘 要: Web 结构优化技术在降低访问延迟、提高访问效率上具有重要作用. 由此提出根据用户访问路径建立 Web 层次模型(WHM), 利用页面相似度合并同层相似页面, 从而构建 Web 概念化模型(WCM). 实验及分析表明, WCM 模型能够在适度聚类的情況下清晰地展现 Web 结构. 此外, 将 WCM 模型应用于预取系统, 该模型所采用的聚类算法在预取效率方面明显优于传统方法, 具有可行性和高效性.

关键词: Markov Tree 模型; Web 概念化模型(WCM); 页面相似度; 聚类

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2015)01-0156-05

A New Method of Web Conceptual Model Based on Web Page Clustering

YAO Yao¹, WANG Zhan-hong², SHI Lei³

(1 School of Information Engineering, Zhengzhou University, Zhengzhou 450044, China;

2 Department of Police Technology, Railway Police College, Zhengzhou 450053, China;

3 School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: Web structure optimization technique plays a significant role in reducing latency and improving the access efficiency. This paper presents the Web Hierarchy Model (WHM) which is constructed based on user access path. Then the Web Conceptual Mode (WCM) is obtained by clustering pages in the same hierarchy according to the pages' similarity. According to the experiments and analyzes, WCM can show the structure of Web clearly under moderately clustering. Furthermore the model is implemented in web prefetching system, in the aspect of prefetching efficiency, the clustering algorithm used in this model is superior to the traditional method and it is feasible and efficient.

Key words: Markov Tree model; Web Conceptual Model; page similarity; cluster

1 引言

随着网络在广度和深度上的飞速发展, 如何帮助用户快速有效地获得自己期望的信息, 克服网络“迷航”, 提高用户访问效率是一个迫切需要解决的难题. Web 结构优化技术^[1]与预取技术^[2]是解决此问题的有效方法. 两种技术的结合成为目前研究的一个热点. Web 结构优化技术分为基于站点模型和基于用户行为两大类. 后者应用更为广泛, 大多数

Web 结构优化系统又是基于 Web 日志^[3]的, 如 Web Utilization Miner、WebLogMiner、Speed Tracer、WebSIFT 和 Web 个性化推荐系统等.

本文研究基于 Web 日志挖掘的优化方法. 通过挖掘用户访问路径建立初步的网站层次结构, 随着层次递增, 页面数量大幅度增加, 可依据同层页面的相似性进一步优化网站层次结构, 从而降低访问代价, 节省网络开销, 也益于提升网站的自身质量.

1998 年, Durand 等人^[4]第一次提出 MAPA 系

收稿日期: 2014-02-23; 修回日期: 2014-04-17

基金项目: 国家自然科学基金项目(U1304603); 河南省教育厅科学技术研究重点项目(13A520651)

统,通过最短权重路径提取层次化导航,但会导致潜在的图存在,破坏了 Web 网站的概念化层次结构。2011 年 Lin 等人^[5]提出基于超链接和网络拓扑结构挖掘 Web 结构,难以反映用户浏览时所处的当前位置。贺秋芳等人^[6]提出将网页内容信息相似度和链接关系结合起来进行聚类,从而识别网站结构,对语义和文本内容的分析又限制了以上方法的应用。Li 等人^[7]提出基于页面检索的页面分类技术构建 Web 层次导航路径(HNP),但忽略了 HNP 的复杂度问题。张春娜等人^[8]提出基于访问路径的 K-PathSearch 算法实现 Web 站点的结构优化,缺乏对站点自身层次结构的考虑。

以上研究对挖掘 Web 网站的层次结构起到了积极作用,但由于尚未考虑 Web 层次和页面之间相似性的协同问题,从而导致相应研究结果不甚理想.为研究上述问题,本文基于页面之间相似性为出发点,基于用户访问路径建立 Web 层次化模型(Web Hierarchy Model, WHM);然后改进一种页面聚类算法,对相似的同类页面聚类,从而建立 Web 站点的概念化模型结构(Web Conceptual Model, WCM).最后,采用模拟日志实验,验证了方法的有效性.

2 基于 Markov Tree 模型提取 Web 层次化模型 (WHM)

2.1 基于 Markov Tree 提取转移概率矩阵

本文采用 Markov 过程模拟用户对网站的访问过程,即是对一组网页的请求序列. Markov 模型表示为三元组 $\langle T, S, P \rangle$,其中 T 是可被用户执行的所有可能操作的集合; S 是 Markov 模型建立的所有状态的集合; P 是一个 $|S| \times |T|$ 的转移概率矩阵.

模拟用户浏览行为的 Markov 链是各态历经的 (ergodic), 因为它可以经过一步或多步转移, 从某一个状态转移到其他状态. 给定用户访问页面的请求序列 S , 构建 Markov 树模型, 结点信息包括 Data 域、SelfCount 域、ChildNum 域和 ChildCounts 域, 具体说明如表 1 所示.

表 1 Markov 树结点信息参数说明

域名称	参数说明
Data	代表当前结点
SelfCount	用于记录结点在上下文出现的次数
ChildNum	表示其孩子结点的个数
ChildCounts	表示结点所有孩子结点出现的次数和

构造 Markov 树的过程如算法 1 所示.

算法 1:构造 Markov 模式树

输入: 用户访问序列 S

输出：Markov 预测模式树

```

1.  $T := \text{Null}$ ,  $i = 0$ ,  $\text{MaxDepth} = 2$ ;
2. while  $i = \min(\text{S\_length}, \text{MaxDepthTree})$ 
3. {  $\text{ss} = \text{Right}(S, i)$ ,  $p = T$ ;
4.   if  $\text{ss\_length} = 0$ 
5.      $p.\text{SelfCount}++$ ;
6.   else
7.     {For  $j$  in  $\text{ss}$  do
8.       {if  $j \neq \text{Last}(\text{ss})$ 
9.          $p = \text{CurrNode}$ ;
10.      else
11.        { $p.\text{ChildCounts}++$ ;
12.         if  $p$  不存在  $\text{ChildNode}$ 
13.           {添加新结点  $\text{CurrNode}$ ;
14.             $p.\text{NumChildren}++$ ;}
15.          $p = \text{CurrNode}$ ;
16.         if  $p = \text{Last}(\text{ss})$ 
17.            $p.\text{SelfCount}++$ ;} } }
18.      $i++$ ;}
19. Return  $T$ .

```

若给定一个包含 4 个会话的访问序列“ $S_1 S_2 S_3 S_1 S_5 S_6 S_1 S_5 S_8; S_1 S_2 S_7 S_4; S_1 S_2 S_7 S_3 S_1 S_2 S_7$ ”, 其 Markov 模式树如图 1 所示, 各结点内的数字分别表示 SelfCount、ChildNum 和 ChildCounts.

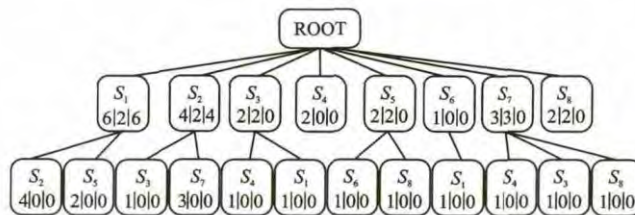


图 1 给定序列且深度为 2 的 Markov 模式树

根据 Markov 模式树计算初始转移概率为每一个结点的 SelfCount 与其双亲结点的 ChildCounts 的比值. 增设开始和结束结点, 即可生成转移概率矩阵, 该矩阵是稀疏矩阵, 结点间的权值如表 2 所示.

2.2 提取 Web 层次化模型 (WHM)

提取 WHM 模型的主要步骤如下所示.

第一阶段:选择种子结点. 由于 Start 页面的初始度最大, 可选 Start 结点为种子结点.

表 2 转移概率矩阵的三元组

Row i	Column j	value
0	1	1.0
1	2	0.67
1	3	0.33
2	3	0.25
2	7	0.75
3	1	0.5
3	4	0.5
4	9	1.0
5	6	0.5
5	8	0.5
6	1	1.0
7	3	0.33
7	4	0.33
7	8	0.33
8	9	1.0
9	1	1.0

$$P = \text{Node}(\text{Max}\{\sum_{j=1}^n p_{i,j}\}) \quad (1)$$

第二阶段: 选择候选双亲结点集. 对每个结点 S_i , 查找其所有的入链 $p_i (i = 1 \cdots m)$, 构成其候选双亲结点集 $\text{Par}_i = \{P_{P1}, P_{P2}, \cdots P_{Pm}\}$, 其中 m 为所有指向当前结点 S_i 的页面数量和.

第三阶段: 确定主双亲结点. 选择 P_i 的候选双亲结点集中第一次出现的最频繁公共候选双亲结点为其主双亲 $F(p_i)$, 创建树结点.

$$F(p_i) = \text{Node}(\text{Max}\{\sum_{p_j \in P} p_{i,j}\}) \quad (2)$$

第四阶段: 确定层次, 删除分支链接. 当前结点的层次等于其主双亲结点所在层次加 1. 主干链接体现在层与层之间的联系, 其他则是分支链接. 示例会话的网站 WHM 模型如图 2 所示, 虚线指向表示为当前结点的主双亲.

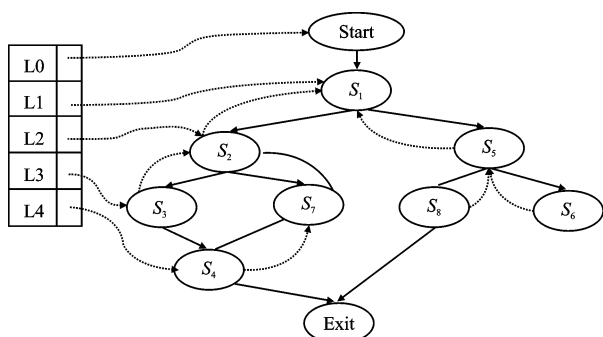


图 2 Web 层次化模型图

3 基于聚类的 Web 概念化模型 (WCM)

为了进一步优化 WHM, 采取根据页面相关性聚类同层网页的方法. 用户对页面访问的频率越高, 说明用户对页面越感兴趣, 并且兴趣背景相似的用

户所访问的页面也越相关. 聚类的关键在于如何查找相似的页面, 结合页面的两种链接方式——内链和外链, 本文给出一种基于行相似和列相似的计算方法来合并相似页面.

3.1 相似性的判定

本文使用内链 / 外链相似度衡量页面之间的相似性, 若相似度小于阈值则判定为相似. 测量时对两个页面的内 (外) 链赋予不同的权重以反映它们对不同层次的重要程度, 记作 $w_{i,j,l}$.

定义 1 外链相似度 (列相似度):

$$\text{OutSim}_{i,j}^2 = \frac{\sum_{l=1}^N w_{i,j,l} \cdot (q_{i,l} - q_{j,l})^2}{\sum_{l=1}^N w_{i,j,l}} \quad (3)$$

$$\text{式中}, w_{i,j,l} = \begin{cases} 0, & q_{i,l} = q_{j,l} = 0 \\ 0.5, & q_{i,l} \neq 0 \text{ 且 } q_{j,l} \neq 0. \\ 1.5, & \text{其他} \end{cases}$$

定义 2 内链相似度 (行相似度):

$$\text{InSim}_{i,j}^2 = \frac{\sum_{l=1}^N w_{i,j,l} \cdot (q_{l,i} - q_{l,j})^2}{\sum_{l=1}^N w_{i,j,l}} \quad (4)$$

$$\text{式中}, w_{i,j,l} = \begin{cases} 0, & q_{l,i} = q_{l,j} = 0 \\ 0.5, & q_{l,i} \neq 0 \text{ 且 } q_{l,j} \neq 0. \\ 1.5, & \text{其他} \end{cases}$$

根据式 (3) 和式 (4) 计算外链和内链相似度, 取二者平均值作为平均相似度.

定义 3 平均相似度:

$$\text{AvgSim}_{i,j} = \frac{\text{OutSim}_{i,j} + \text{InSim}_{i,j}}{2} \quad (5)$$

矩阵中的元素值代表相似度, 其值越接近 0, 说明相似度越高. 从矩阵角度出发, 外链相似即列相似, 内链相似即行相似. 设定一个阈值 ϵ , 当相似度小于阈值时说明两个结点相似. 理论上阈值越大, 相似的状态越多, 但是聚类准确率就会下降. 合适的阈值能够在尽量不丢失状态转移行为的同时把 WHM 模型聚类到一个利于有效预测的大小, 实验中阈值分别设定为 0.1 和 0.15.

聚类时以行主序方式扫描平均相似度矩阵, 对位于同一层次且其相似度小于给定阈值的结点进行合并操作生成新结点, 转移概率矩阵的计算公式如下^[9]:

$$\begin{aligned} P_{k,i \vee j} &= P_{k,i} + P_{k,j} \\ P_{i \vee j,k} &= \frac{P_{i,k} \cdot \sum_k P_{k,i} + P_{j,k} \cdot \sum_k P_{k,j}}{\sum_k P_{k,i} + \sum_k P_{k,j}} \end{aligned} \quad (6)$$

3.2 聚类算法描述

算法 2 聚类算法

输入: 用户访问序列 S 的 Markov Tree 模型.

输出: 聚类后的转移概率矩阵 P_a .

1. 计算初始转移概率矩阵 P_0 ;
2. 行主序扫描 P_0 , 根据式(3)、(4)、(5) 计算外链相似度矩阵、内链相似度矩阵以及平均相似度 AveSim Matrix;
3. For WHM 层次 $h=H$ down to 1
4. For AveSim Matrix 矩阵的每一行
5. For AveSim Matrix 矩阵的每一列
6. If AveSim Matrix $(i, j) < \epsilon$ and $\text{OutSim}_{i,j} < \epsilon$ and $\text{InSim}_{i,j} < \epsilon$
7. 根据式(6)聚类相似结点;
8. 输出聚类后的转移概率矩阵 P_a .

3.3 构建 WCM 模型

根据聚类后得到的转移概率矩阵提取出其 Web 概念化模型结构. 第一层代表主页, 其他层次的结点由聚类和非聚类页面构成. 图 2 相对应的 WCM 模型如图 3 所示.

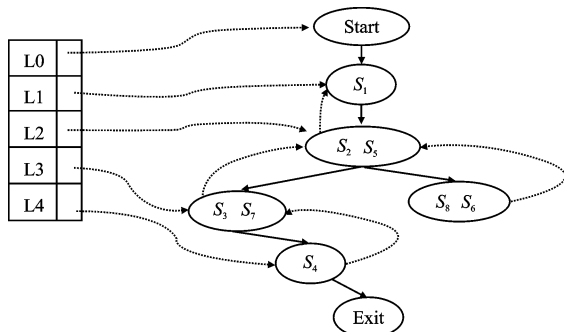


图 3 聚类后的 Web 概念化模型

4 实验设计与分析

4.1 实验设计

实验依据真实日志文件, 采用 Markov 方法挖掘用户路径, 在此基础上进一步构建 WHM 模型和 WCM 模型, 最后应用于 Web 预取系统, 衡量其预取效果.

实验结果评价指标分别是预取准确率、查全率^[10]、预取效率以及聚类比率. 其中 PRS 预取效率是基于准确率和查全率为参数的, 综合衡量预取效果. 计算公式如下:

$$PRS = \frac{2 * Pr * Re}{Pr + Re} \\ = \frac{2 * (P^+ / (P^+ + P^-)) * (R^+ / |R|)}{P^+ / (P^+ + P^-) + R^+ / |R|} \quad (7)$$

式中, Pr 代表准确率, Re 代表查全率, P^+ 和 P^- 分别

表示正确的和不正确的预测数, R^+ 表示模型预测的请求数, $|R|$ 表示总请求数.

4.2 实验结果

实验一选用 Berkley 大学计算机系实验室^[11] 日志文件为数据源. 初始日志包含 12 102 条记录; 数据清理后得到 1 981 条记录. 相似度阈值设为 0.1. Web 概念化建模聚类效果如表 3 所示, 表中的链接代表指向当前层的链接数目.

表 3 模型聚类效果

层次	WHM		WCM		聚类率/%	
	结点	链接	结点	链接	结点	链接
1	1	0	1	0	0	0
2	19	19	3	3	84.2	82.1
3	145	176	67	67	53.8	62.2
4	432	478	256	287	40.7	39.9
5	231	255	167	198	27.7	22.5
6	78	85	63	68	19.2	19.9
7	67	72	55	61	17.9	15.0
8	34	37	30	34	11.8	9.0
9	24	24	23	24	4.2	2.0
10	10	10	10	10	0	0

对比 WHM 和 WCM 模型, WCM 结点数和链接数均大幅度减少, 位于中间层的页面数量多, 存在相似性的可能性较大, 故聚类比例较大. 随着层次继续递增, 用户浏览的页面数量下降, 相似页面也相对减少.

实验二选用信息网络重点实验室^[12] 日志文件为数据源, 数据清洗后, 随机选取 4 组数据集, 非 0 元素所占比例为稀疏度呈依次递减, 相似度阈值分别设为 0.15 和 0.1, 分析 WHM 模型和 WCM 模型的预取效果, 结果如表 4 所示.

表 4 预取效果比较

数据集	稀疏度	WHM			WCM		
		Pr	Re	PRS	Pr	Re	PRS
1	5.6	46.2	56.5	50.8	37.5	61.5	46.6
2	6.2	55.5	47.3	51	53.8	56.5	55.2
3	8.75	53.8	56.5	55.2	50	58.3	53.8
4	12.3	66.7	47.4	55.4	70	50	58.4

根据实验二的结果可知随着数据集提取的转移概率矩阵的数据密集度增大, PRS 指标有所升高. 对比聚类前后, 模型的预测准确率仅受到一点影响, 分别下降了 8.7%、1.7% 和 3.8%, 另外一组数据显示提高了 3.3%. 而查全率均有相应的提高, 分别是 5%、9.2%、1.8% 和 2.6%. 对 PRS 的影响很小, 在 4% 左右. 对于一些数据, 可能聚类掉的页面恰好是无用的页面, 此时就会出现准确率提高的现象. 聚类的过程使得状态的特性趋于一般化, 故查全率均有

提升.

实验三基于实验二的日志,四个数据集的数据密度依次递增,此时对比本文聚类方法和传统聚类方法(K-means 和 PageGather)对预取性能(PRS)的影响,如图 4 所示.

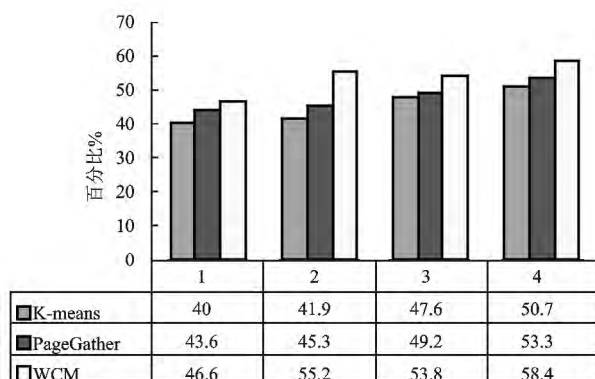


图 4 不同聚类算法对预测模型的综合影响

整体上,数据越密级,页面相似的可能性越大,三种方法的预取性能越有优势.当数据密集度低于 10%时,K-means 和 PageGather 算法对于数据的稀疏性并不敏感,预取效率维持在不高的水平.而 WCM 模型采用的聚类算法相比传统方法在预取效率方面有明显的优势,具有可行性和高效性.

5 结束语

基于用户对 Web 站点的访问偏好,本文提出了 WCM 建模方法,它是一种 Web 站点结构优化方法,用于聚类 Web 层次模型中同层次的相似性页面.文中首先考察基于 Markov Tree 模型提取 WHM 结构,具体地引入主双亲结点确定结点所在层次.再者,随着层次递增,页面数量大幅度增大,根据页面内(外)链相似度对同层页面的相似状态合并,优化 WHM 模型.最后,应用 WCM 模型于预取系统.实验验证了该模型在预取性能和聚类效果方面表现的优势,对 Web 预取技术和 Web 结构优化技术的结合和应用具有很好的参考价值.

参考文献:

[1] 李亮. 基于 WEB 挖掘的网站结构优化技术研究[D].

西安:西北师范大学,2012.

- [2] 班志杰,古志民,金瑜. Web 预取技术综述[J]. 计算机研究与发展, 2009,46(2):202-210.
- [3] 万常选,邓松,刘喜平,等. Web 数据源选择技术[J]. 软件学报,2013,24(4):781-797.
- [4] Durand D G, Kahn P. MAPA: A system for inducing and visualizing hierarchy in websites[C]// Proceedings of ACM Hypertext '98. New York: ACM, 1998: 66-76.
- [5] ShianHua Lin, KuanPak Chu, ChunMing Chiu. Automatic sitemaps generation: Exploring website structures using block extraction and hyperlink analysis [J]. Expert Systems with Applications, 2011, 38(4): 3944-3958.
- [6] 贺秋芳,曾后杰,蔡延光. 挖掘用户标签的增强型社区网页聚类算法[J]. 微电子学与计算机,2013,30(2): 74-77.
- [7] Jian-Qiang Li, Yu Zhao, Hector Garcia-Molina. A path-based approach for web page retrieval [J]. World Wide Web, 2012, 15(3):257-283.
- [8] 张春娜,李铁然. 基于路径聚类的页面访问次序的挖掘[J]. 计算机工程与设计,2013,34(1):303-306.
- [9] LeiShi, Yao Yao. On the compression of markov prediction model[C]// Proceedings of Fourth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE Computer Society Press. Haikou, China: IEEE,2007:512-516.
- [10] B Wu, A D Kshemkalyani. Objective-optimal algorithms for long-term Web prefetching [J]. IEEE Transactions on Computers, 2006, 55(1):2-17.
- [11] Computer Science Department, UC, Berkeley. [EB/OL]. [2013-02-23]. <http://www.cs.berkeley.edu/logs>.
- [12] Henan Provincial Key Lab on Information Network. [EB/OL]. [2013-02-23]. <http://netlab.zzu.edu.cn>.

作者简介:

姚 瑶 女,(1982-),硕士,讲师.研究方向为网络信息技术和 Web 挖掘. E-mail:wmyy@163.com.

王战红 男,(1979-),硕士,讲师.研究方向为网络信息技术.

石 磊 男,(1967-),博士,教授.研究方向为高性能计算、Web 预取与挖掘技术等.