计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
42(5):765~770, 2005

# 根据用户行为网上导航的方法

杨 捷[1] 毋国庆[2]
[1](华南理工大学计算机软件学院 广州 510006)
[2](武汉大学计算机学院 武汉 430072)
(yjclear@scut.edu.cn)

# Associated Navigation on the Web According to Users' Activities

Yang Jie[1] and Wu Guoqing[2]
[1](*School of Software Engineering of South China University of Technology*, *Guangzhou* 510640)
[2](*School of Computer Science*, *Wuhan University*, *Wuhan* 430072)

**Abstract** With the growth of the Internet, World Wide Web increasingly helps people to make good use of rich information from local or remote site. The amount of Web pages is so enormous that users are destined to drown in the huge data of the Web without any navigation. To provide a new navigation approach, a modified Markov chain model is introduced, which utilizes all group members' traces in the Web to recommend some potential useful Web sites and navigates people when they browse Web pages, while users' activities react to the model. Before that, an algorithm based on a semi-formal description of process is necessarily given for collecting desired data to gain top grade results. The method is also illustrated by analyzing the proxy server's access log in the prototype system.

**Key words** data mining; Web navigation; data collecting; Markov chain

**摘 要** 随着因特网的成长,网络浏览使人们从本地或远程更方便地获取各种信息.网页数量的疯狂增长已经使得用户面对庞大的数据群无所适从,急需导航技术的帮助.一个新的马尔可夫链模型被引入用来跟踪所有团体成员的网页访问活动,并且推荐一些有用站点,引导人们更有效率地浏览网站.还提出一个基于半形式化过程描述的数据搜集算法,来获得有用数据,以推导出最好结果,并在原型系统中分析了代理服务器上的访问日志,对该算法进行描述.

**关键词** 数据挖掘;网络导航;数据搜集;马尔可夫链

**中图法分类号** TP311

## 1 引 言

网络时代的互联网越来越重要,通过它人们可以方便、迅速地获取更多信息.但与此同时,网页的数量也在飞快增长,以至于现在"如何找到有效方法,从网页中挖掘有用数据"已经成为当务之急.现在已有一些导航技术帮助人们更好地浏览网页,例如大家熟知的搜索引擎,给定一个关键词它将快速返回所需的结果,但遗憾的是大部分情况下人们无法给引擎足够详细和确切的合适信息,所以无可避免地,如果给的要求限制太松,返回结果会包括很多无关信息;反之如果限制太严格会错失一些重要信息.另外一种导航技术是某种浏览器的书签功能

收稿日期:2004-01-07;修回日期:2004-10-21
基金项目:国家自然科学基金项目(69873035);华南理工大学自然科学基金项目(G03-E5041450)

(如 Internet Explorer,Netscape 等),它可以帮你在浏览器中自如保存喜爱的网页(内容和地址),问题是浏览器只能保留网页供以后使用,而不能向人们推荐潜在的有用网页,即它是被动的,而非主动. 许多研究致力于通过一些新方法来弥补传统技术的缺陷,以改善这种不尽人意的状况,其中最常见的是网页中的数据挖掘,即网页挖掘. 根据挖掘的对象网页挖掘被分为 3 种[1]:网页内容挖掘、网页结构挖掘、网页用法挖掘. 网页内容挖掘根据网页内容向用户提供有用信息,但是由于网页数据千奇百怪而且是非结构化的,所以很难奏效;网页结构挖掘的目的是找到以超链接拓扑结构为基础的网络链接结构的基本模型,虽然也有些令人振奋的结果,但我们还是认为网络组织结构的紊乱使得这项工作十分艰难;网络用法关注用户浏览网页时从用户行为得到的数据(包括 3 类日志:网络服务器数据、代理服务器数据、客户数据)和其他用户数据. 基于此,我们的主要目标是开发一种网页挖掘的更高效的途径. 给出一个模型分析代理服务器的日志数据,并预言一些潜在的受欢迎的网页. 日志数据直接反映用户的活动,但因为我们可以脱机分析日志数据,所以不会影响代理服务器的性能. 根据文献[2],我们有 3 种日志数据可选. 服务器日志和客户日志分别特别强调服务器的活动和特定的用户活动,相反作为和外界网络通信的惟一方法,代理服务器记录了每个人的每个浏览网页的活动,并大大帮我们预言“热”页.

我们在第 2 节中给出了基于马尔可夫链的增强模型;第 3 节在模型基础上提供了数据收集和挖掘、并以此进行用户网络导航的方法;第 4 节给出整个系统的体系结构原型;在第 5 节使用实验数据对该方法进行评估;第 6 节是结论.

## 2 新的马尔可夫链模型

有多种方法用于团体导航,例如马尔可夫链、决策树、贝叶斯网、熵、聚类等. 实际上相当多的时候用户会浏览他们刚浏览过的网页甚至重复整个浏览历史,马尔可夫链模型[3]是一种非常经典的概率模型,适合于无记忆情形(下一步状态仅仅依赖前一步),所以我们在众多方法中选择它. 另外由于下一网页可能会依赖许多历史页,必须加一个参数到模型中,表示做决策时要考虑多少天的数据,这样就引入了动态转换矩阵以弥补原马尔可夫模型的缺陷.

**定义 1.** 设 $M$ 是一个 $k \times k$ 矩阵,有元素 $m_{i,j}(i,j)=1,2,\cdots,k$. $m_{i,j}$表示如果前一状态是 $s_i$(其中 $s_i$ 表示用户浏览第 $i$ 个网页,假定网页已经被按顺序编号),那么下一状态为 $s_j$(用户浏览第 $i$ 个网页)的概率将为 $m_{i,j}$(其中有 $1 \geqslant m_{i,j} \geqslant 0$). 一个有限空间 $S=(s_1,s_2,\cdots,s_k)$的随机过程$(X_0,X_1,\cdots)$被称为一个有转换矩阵 $M$ 的马尔可夫链,如果对所有的 $i,j \in \{1,2,\cdots,k\}$,且所有 $i_0,\cdots,i_{n-1} \in \{1,2,\cdots,k\}$均有

$$P(X_{n+1}=s_j \mid X_0=s_{i_0},X_1=s_{i_1},\cdots,X_{n-1}=s_{i_{n-1}},X_n=s_i)=P(X_{n+1}=s_j \mid X_n=s_i)=m_{i,j},$$

转换矩阵 $M$ 必须满足下面两个条件:

$$m_{i,j} \geqslant 0 \text{ for all } i,j \in \{1,\cdots,k\}, \tag{1}$$

$$\sum_{j=1}^{k} m_{i,j}=1 \text{ for all } i \in \{1,\cdots,k\}. \tag{2}$$

**定义 2.** 设 $U$ 为一个向量,维数是状态(网页)数;$(u_1,u_2,\cdots,u_k)$是概率分布$(u_i=P(X=s_i))$;$U^{\text{prev}},U^{\text{next}}$各表示前和后的概率分布;$X$ 代表一个随机过程,$s_i$ 表示第 $i$ 个状态,则有两个性质:

$$\sum_{i=1}^{k} u_i=1, \tag{3}$$

$$U^{\text{next}}=U^{\text{prev}} \times M. \tag{4}$$

虽然马尔可夫链模型善于描述随机事件,却不能描述状态数或者转换矩阵易变的情况,因此给出如下定义.

**定义 3.** 设 $k$ 为一个变量(其值依赖于最近数据),$M$ 是一个 $k \times k$ 矩阵,其中每个元素记为 $m_{i,j}$ $(i,j=1,\cdots,k)$;随机过程$(X,X')$在修改后的模型中有状态空间 $S=(s_1,s_2,\cdots,s_k)$. 可知:

$$P(X'=s_j \mid X=s_i)=m_{i,j}. \tag{5}$$

**定义 4.** 设 $L$ 为一个 $k$ 维向量,元素 $l_i$ $(i=1,2,\cdots,k)$. 如果对所有的 $i \in \{1,2,\cdots,k\}$,有 $l_i \geqslant 0$,那么 $L$ 被称为一个数据记录. $L^n$,$L^{\text{new}}$分别表示第 $n$ 天和最近的数据.

为了重生成转换矩阵,需使用最近数据和历史数据重新计算所有矩阵中的元素[4]. 这里有两个新的参数 $\alpha$(表示把哪一天的数据当成最近数据),$\gamma$(根据最近使用的频率将数据分级)[5]. 如 Simon 所说,数据的权将几何级数递减. 当 $\alpha$ 的值为 $m$ 且 $L^n$一个给定网页在第 $n$ 天被访问的次数,$L^{\text{new}}$是历史数据的加权之和:

$$L^{\text{new}}=\sum_{i=n-m+1}^{n} l^i \times \gamma^{n-1}. \tag{6}$$

显然 $\alpha$ 值为 1 时,新模型将恢复为原马尔可夫

模型.此外当 $n$ 个网页最近数据为($l_1,l_2,\cdots,l_n$)($l_k>0$)时,得到新的转换矩阵 $\boldsymbol{M}$,其中,

$$m_{i,j}=\begin{cases}l_j\Big/\sum_{k=1,k\neq i}^{n}l_k, & i\neq j,\\ 0, & i=j.\end{cases}\tag{7}$$

**定理1.** 设 $L=(l_1,l_2,\cdots,l_n)$,其中,$l_k>0(1\leqslant k\leqslant n)$,则由式(7)决定的 $\boldsymbol{M}_{n\times n}=(m_{i,j})$ 满足式(1),(2).

**证明.**

(1) 由 $l_1,l_2,\cdots,l_n>0$,我们有 $m_{i,j}\geqslant 0$,因此 $\boldsymbol{M}_{n\times n}=(m_{i,j})$ 满足式(1).

(2) $\sum_{j=1}^{n}m_{k,j}=\sum_{j=1,j\neq k}^{n}m_{k,j}+m_{k,k}=\sum_{j=1,j\neq k}^{n}\left(l_j\Big/\sum_{l=1,j\neq k}^{n}l_l\right)+0=\left(\sum_{j=1,j\neq k}^{n}l_j\right)\Big/\left(\sum_{l=1,l\neq k}^{n}l_l\right)=1$,因此,$\boldsymbol{M}_{n\times n}=(m_{i,j})$ 满足式(2)且证明完整. 证毕.

## 3　新的用户网络导航方法

用户网络导航的方法建立在导航模型的基础之上,主要由数据搜集算法和数据挖掘算法组成[6,7].

### 3.1　数据采集

输入日志数据到我们的模型之前必须对它们预处理.因为在代理服务器中(如 Jigsaw,Squid),日志是某种冗余信息,会降低系统性能、耽误处理时间、产生某种不确定性[8];我们必须按利益整理日志以避免数据集中出现重复项,并给每个项一个相应的逐日访问计数.过于详细的链接地址不满足我们的要求,所以要从日志文件中抽取有用的超级链接[9,10].例如团队成员曾访问雅虎新闻站点两次,其中之一浏览某条新闻而另一个浏览的是同一站点上的另一条,假定这两个用户浏览的地址分别为 http://story.news.yahoo.com/fc?cid=34&tmpl=fc&in=World&cat=Turkey 和 http://news.yahoo.com/fc?tmpl=fc&cid=34&in=business&cat=downsizing_and_layoffs).最后所有对用户无用的"脏数据"如广告(会给计算结果带来恶劣影响)要被过滤出来.

数据是得出决策的基础,必须有一个明确的方法获得最终的数据集,但是很少有文章谈到如何得到这个数据集.在我们的代理服务器中每日访问日志(全部是文本)量大约是 4MB 或 5MB,因此这里引入归纳的方法,该法擅长处理这类有限和可数无限问题[3,11].原始日志数据集 $L$,$T$ 将集合 $R$ 的元素转换到 $L$,记为

$$\forall r\in R\rightarrow T(r)\in L,\tag{8}$$

(如果 $r$ 属于 $R$,一个和 $r$ 相关的 $T(r)$ 属于 $L$)在转换 $T$ 中,从未加工的日志文件中抽象出有用信息,并修改超级链接以供后用.$L$ 中的元素个数不大于 $R$ 中的元素个数,可以证明 $T$ 是一个映射,因此转换可以缩减数据并提高系统性能.通过过滤可以进一步缩减数据集合:

$$\forall r\in R\rightarrow \text{if}(\neg F_1(r)\wedge\neg F_2(r)\ \text{then}\ T(r)\in L,\tag{9}$$

其中,$F_1$ 和 $F_2$ 是两个用于检查是否 $T(r)$ 属于 $L$ 的卫士函数,$F_1(r)$ 的值表示 $r$ 是否应该除去,$F_2(r)$ 表示对 $r$ 同样的判断.数据要通过两个测试才属于 $L$.搜集数据的算法输入为整理过的数据集合,输出的结果作为分析的基础提供给修改过的马尔可夫链模型.

图1所示的算法有两个新的变量:*AddrLen* 是地址的长度由'/'分隔,*AutoDetection* 表示程序是否过滤广告页.$L$ 还包括一个每个 $T(r)$ 在 $L$ 中出现的相关次数的计数.

```
For each (r in R)
  if the date in r is not required by us
    then continue;
  if the commend in r is not "GET"
    then continue;
based on the argument AddrLen to decide from where we should
truncate the http address (we get T(r))
if(AutoDetection != 0)
  then {check whether T(r) is an advertisement page;
    if yes then continue};
if(exists(T(r),L))
  then add one to the counter of T(r)
  else add T(r) to the set L
end foreach
```

Fig. 1　Algorithm for collecting data.

**图1　数据搜集算法**

使用归纳的方法生成我们的最终数据集,还可以证明它的一些属性,例如,$L$ 中没有相同的地址;转换 $T$ 是满射,等等.为简洁起见,这里不给出具体证明.

### 3.2　网络导航

有了数据采集的方法和在第2节中分析设计的增强型马尔可夫链模型,我们的数据挖掘算法(见图2)的输入由 $\alpha$(其中 $\alpha$ 表示历史数据的长度,$\alpha$ 小于 $h$),$\gamma$,今天的数据和相应的历史数据($l^1=(l_1^1,l_2^1,\cdots,l_n^1),\cdots,l^h=(l_1^h,l_2^h,\cdots,l_n^h)$)组成.输出是 $U=$

$(u_1,u_2,\cdots,u_n)$,其中 $u_i=P(X'=s_i)$(用户下次将浏览网页 $s_i$ 的概率). 如果需要也可以引入一个起始页 $l^{\text{init}}$作为一个一维矩阵 {}.

```
for i = 1 to n
    l_i^new = 0
    for j = 1 to α
        l_i^new = l_i^new + l_i^(h-j+1) * γ^(j-1)
    end (for j)
end (for i)
sum = Σ_{i=1}^{n} l_i^new
for i = 1 to n
    for j = 1 to n
        if i = j then {m_{i,j} = 0 ;continue}
        m_{i,j} = l_j/(sum - l_i)
    end (for j)
end (for i)
l^output = l^new * M_{n×n}(m_{i,j})
U^output = (u_1, u_2, ..., u_n), where
u_i = l_i^output / (Σ_{j=1}^{n} l_j^output)
```

Fig. 2 Algorithm for data mining.

图 2 数据挖掘算法

## 4 原型系统

在历史数据的基础上配置完后,使用我们修改过的马尔可夫链模型进行自动计算后给出一个排行榜.系统引导用户时,用户的活动也影响系统的动作,如果用户遵从我们的引导,有用的网页将会被包括在榜单中. 通过和代理服务器的交互,用户浏览活动的不确定性将降低,目的性将加强,原型系统的体系结构如图 3 所示. 日志数据输入到模型后,我们可以通过计算得到一个关于下一个网页的预言,该预言有两种提交形式:排行榜将被放到代理服务器上并自动提供给用户;开始页应该作为用户默认页(作为对网页的入口提供服务)被生成并放到Web服务器上. 我们的模型将根据来自用户的反馈重新调整计算,作为简单系统中的重要特征,输入数据是关于一个团体或者一个实验室的而不是单个人的,输出将反馈到我们的模型并推导出新的结果[6,12].

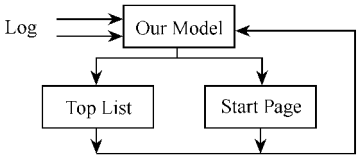


Fig. 3 Architecture of our system.

图 3 原型系统

## 5 实 验

在应用我们的原型系统评估模型时,应关注以下方面:它是否能得到满意的结果;程序是否能高性能运行. 我们的实验环境为 4CPU(296MHz)计算机,2GB RAM,操作系统 Sun OS 5.6. 从代理服务器上随机下载连续 10 天的日志文件用于检验本网页导航方法,并且通过修改后编译代理服务器软件的源代码来实现新方法. 系统分为两部分(见图 4),其中一部分负责搜集数据、分析数据并提供排行榜(用 Perl 写成);另一部分提交榜单给用户并显示每个超级链接的访问次数. 表 1 中列举了需要解释的几个参数意义. 在接下来的图 5 至图 6 中给出了结果.

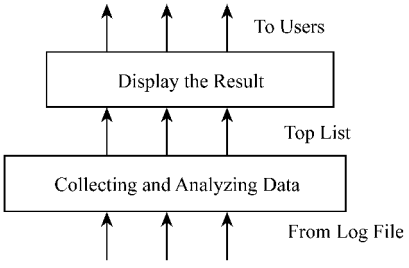


Fig. 4 Structure of our system.

图 4 系统结构

**Table 1 Parameters in Our System**

表 1 系统参数说明

| AddrLen | HistoryDays ($\alpha$) | Affection ($\gamma$) | AutoDetection |
|---|---|---|---|
| From where to obstruct the hyperlinks | Log data of how many days used for analysis | The factor to reduce the affection of history data | Whether filtering hyperlinks automatically |

在每个图中都有两个曲线:排行榜中榜数(今天在我们的预言中多少超级链接受欢迎)和全部数据的中榜数(今天的访问数). 图 3 说明,给定一个固定的 $\gamma(=0.9)$,$\alpha$ 越大中榜率越高,即历史数据对预言有用. 换句话说,用户的浏览活动有持续性,反映为历史数据和将来的网页之间的联系.

同样,图 5 至图 8 表示:对一个固定的 $\alpha$(分别为 2,3,4),$\gamma$ 越小中榜率越低. 如果我们减小 $\gamma$(减少历史数据的影响),中榜数就下降. 完整的实验数据见表 2.

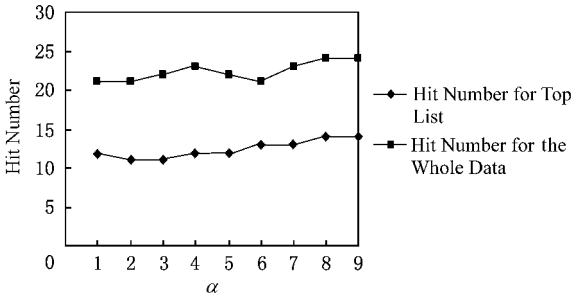


Fig. 5　Hit number of 30 days when $\gamma$ is 0.9.

**图 5　当 $\gamma$ 为 0.9 时预测网页 30 天中的中榜情况**

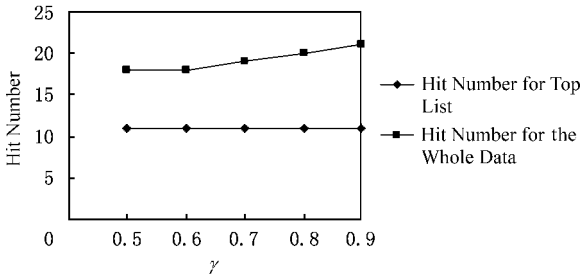


Fig. 6　Hit number of 30 days when $\alpha$ is 2.

**图 6　当 $\alpha$ 为 2 时预测网页 30 天中的中榜情况**

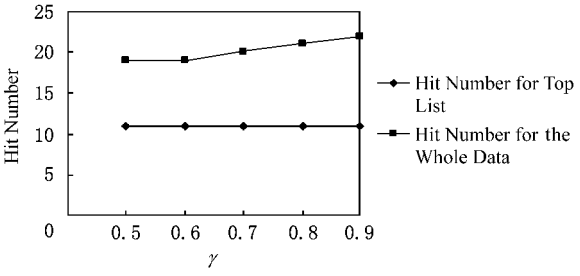


Fig. 7　Hit number of 30 days when $\alpha$ is 3.

**图 7　当 $\alpha$ 为 3 时预测网页 30 天中的中榜情况**

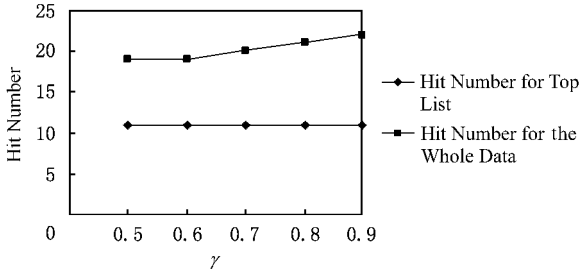


Fig. 8　Hit number of 30 days when $\alpha$ is 4.

**图 8　当 $\alpha$ 为 4 时预测网页 30 天中的中榜情况**

**Table 2　All Experiment Data**

**表 2　所有的实验数据**

| $(\alpha,\gamma)$ | $n=30$ | $N=20$ | $n=10$ | $(\alpha,\gamma)$ | $n=30$ | $n=20$ | $n=10$ |
|---|---|---|---|---|---|---|---|
| (1,*) | 12/21 | 7/11 | 5/8 | (4,0.9) | 12/23 | 9/15 | 3/10 |
| (2,0.9) | 11/21 | 8/15 | 5/8 | (4,0.8) | 12/21 | 9/14 | 5/10 |
| (2,0.8) | 11/20 | 8/15 | 5/8 | (4,0.7) | 12/21 | 9/14 | 5/10 |
| (2,0.7) | 11/19 | 8/15 | 5/8 | (4,0.6) | 11/18 | 9/14 | 5/10 |
| (2,0.6) | 11/18 | 8/14 | 5/8 | (4,0.5) | 11/18 | 9/15 | 5/9 |
| (2,0.5) | 11/18 | 8/13 | 5/8 | (5,0.9) | 12/22 | 9/15 | 3/9 |
| (3,0.9) | 11/22 | 9/15 | 4/9 | (6,0.9) | 13/21 | 10/17 | 4/9 |
| (3,0.8) | 11/21 | 9/15 | 5/10 | (7,0.9) | 13/23 | 10/16 | 4/9 |
| (3,0.7) | 11/20 | 9/14 | 5/10 | (8,0.9) | 14/24 | 10/18 | 4/9 |
| (3,0.6) | 11/18 | 9/14 | 5/9 | (9,0.9) | 14/24 | 10/19 | 4/9 |
| (3,0.5) | 11/18 | 9/15 | 5/9 | | | | |

最后进行系统性能分析. 事实上代理服务器的负担太重很容易成为网络的瓶颈,所以我们将最耗时的工作(使用组织好的数据计算预言)脱机完成. 有一个大的 $\alpha$,要处理的数据也很庞大,处理时间很长. 关于时间的实验数据见表 3 所示：

**Table 3　Experiment Data About Processing Time**

**表 3　处理时间的实验数据**

| $n$ | time | $n$ | time | $n$ | time |
|---|---|---|---|---|---|
| 1 | 23″ | 4 | 1′42″ | 7 | 4′06″ |
| 2 | 40″ | 5 | 2′21″ | 8 | 5′38″ |
| 3 | 1′18″ | 6 | 2′50″ | 9 | 6′53″ |

## 6　结　　论

我们提出了一个改进的马尔可夫链模型,给用户提供一些预言以引导他们更有效地使用网页. 模型使用的是一个团组的数据而不是个别的数据作为输入,并且不仅利用历史数据也利用人们当前的活动以获得结果. 考虑到搜集数据的重要性,本文还用一个具体算法半形式化地描述了如何生成数据集.

为了评估模型,我们分析了代理服务器的访问日志,结果表明我们的模型是有效的、可行的.

**致谢** 本文在京都大学上林实验室完成,非常感谢上林教授和成凯的帮助. 感谢刘珊完成原型系统.

## 参 考 文 献


1 R. Kosala, H. Blockeel. Web mining research: A survey. SIGKDD Explorations. http://www.cs.kuleuven.ac.be/~dtai/publications/files/33042.ps.gz, 2000

2 J. Srivastava, R. Cooley, M. Deshpande, *et al*. Web usage mining: Discovery and applications of usage patterns from Web data. SIGKDD Explorations. http://www.umn.edu/research/websift/papers/sigkddpp.ps, 2000

3 Olle H * ggstr * m. Finite Markov Chains and Algorithmic Applications. Cambridge: Cambridge University Press. http://www.cambridge.org/catalogue/catalogue.asp?isbn=0521813573

4 E. Watson, Y. Shi, Y. Chen. The Web user access modeling and cache performance analysis. Decision Support Systems, 1999, 25(4): 309~338

5 Robert Cooley, Pang-Ning Tan, Jaideep Srivastava. Discovery of interesting usage patterns from web data. University of Minnesota, Technical Report: TR 99-022, 1999

6 Samhaa El-Beltagy, Wendy Hall, David De Roure, *et al*. Linking in context. In: Proc. the 12th ACM Conf. Hypertext and Hypermedia. New York: ACM Press, 2001. 151~160

7 Dieberger, Andreas. Supporting social navigation on the World-Wide Web. Int'l Journal of Human Computer Studies. http://citeseer.ist.psu.edu/dieberger97supporting.html, 1997

8 J. L. Neto, A. D. Santos, C. A. A. Kaestner. Document clustering and text summarization. In: Proc. the 4th Int'l Conf. Practical Applications of Knowledge Discovery and Data Mining. London: The Practical Application Company, 2000. 41~55

9 M. Levene, G. Loizou. Web interaction and the navigation problem in hypertext. Encyclopedia of Microcomputers, 2002, 28(7): 381~398

10 C. Berrut, F. Fourel, M. Mechkour, *et al*. Indexing, navigation and retrieval of multimedia structured documents: The PRIME information retrieval system. ESPRIT III Basic Research Action, Tech. Rep.: 8134(FERMI- D11), 1997

11 N. Zin, M. Levene. Constructing web views from automated navigation sessions. Department of Computer Science, University College London, Tech. Rep.: RN/99/35, 1999

12 J. Touch. The LSAM proxy cache—A multicast distributed virtual cache. In: Proc. the 3rd Int'l WWW Caching Workshop. http://wwwcache.ja.net/events/workshop/14/lsam-arch.pdf, 1998


**Yang Jie**, born in 1977. She is a teacher at the School of Computer Software, South China University of Technology. She received her Ph.D. degree in computer science from Wuhan University in 2004. Her current research areas include formal method and software architecure.

**杨捷**,1977 年生,讲师,主要研究方向为形式化方法、软件体系结构等.

**Wu Guoqing**, born in 1954. He is a professor and doctoral supervisor of the Department of Computer Science, Wuhan University. His current research interests are theory in software, requirement engineering, and auto-programming.

**毋国庆**,1954 年生,教授,博士生导师,主要研究方向为软件理论、需求工程、程序自动化生成等(wgq@whu.edu.cn).

**Research Background**

Supported by the National Science Foundation of China under Grant No.69873035 and the Science Foundation of the South China University of Technology under Grant No. G03—E5041450.Dataminiing is a hot and important research topic of computer science. Despite the growth of Internet and the advances in WWW technology, current methods for web users to make good use of information from so enormous web pages are not as efficient as we would like. The existing Web assisting techniques are usually sorted into search engine (among which the intelligent search engine based on agent is the most notable), bookmark distributed with some browser and some other ones as neuron network. However, none of the traditional navigation methods are satisfactory since the search key always must be precise so austere while the bookmark has passive nature rather than active one. To establish a path towards higher efficiency in web pages mining, a model to analyze the log data of proxy server and forecast some potential popular Web pages is presented. After an algorithm is necessarily given for collecting desired data to gain top grade results, a modified Markov Chain Model is introduced which utilizes all group members' traces to recommend some potential sites and navigates them when browsing Web pages. Based on the history data, our prototype system is active and alternating. It gives a top list (hit parade) to users after calculation using our modified Markov Chain Model, which is fulfilled automatically after configuration. The result is obvious and satisfied. These semi-formal or formal methods can be used in other research areas, too. For example, the property modeling of software architectures in software engineering and the security modeling of the network protocols, etc. Such efforts will be found in our further works.