基于属性重要性的 WUM 数据预处理方式①

王亚军[1]，王传安[1,2]

[1](安徽科技学院，凤阳 233100)
[2](江苏大学 计算机与通信工程学院，镇江 212013)

摘 要：为了降低 Web 日志数据的规模，并能从预处理后的数据中发现更有价值的访问模式，在引入知识的信息量的基础上，给出了单个属性相对于属性集的重要性量化值的概念，并采用了操作系统中 LRU 页面置换算法的思想，提出了基于属性重要性的 WUM 数据预处理方式。实验证明：该方式可以删除不具有挖掘价值的、因用户短期行为而访问的 Web 日志记录，剔除掉噪音数据，从而有效减小了日志挖掘的复杂度。

关键词：访问模式；LRU 页面置换算法；用户短期行为；噪音数据

Data Preprocessing Method Based on Importance of Property for WUM

WANG Ya-Jun[1], WANG Chuan-An[1,2]

[1](Anhui Science and Technology University, Fengyang 233100, China)
[2](College of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: To reduce the Web log data scale and discover more recommendable access patterns from data preprocessed, with knowledge based on amount of information, the concept of quantify value of importance of every property in relation to property set was proposed, and used the idea of LRU page replacement algorithm in the operating system, a new data preprocessing method based on importance of property was proposed. The experiments show that the method could delete Web log records which were caused by user short-behavior and have not mining value, and filter out the noise data. Accordingly it can reduce the complexity of log mining effectively.

Keywords: access patterns; LRU page replacement algorithm; user short-behavior; noise data

1 引言

随着 Internet 的快速发展，网络信息的重要性已经引起人们越来越大的关注，Web 挖掘[1]的目的就是从海量的网络数据中发现潜在的有用模式和隐藏的信息。Web 挖掘可分为:Web 内容挖掘，Web 结构挖掘，Web 使用挖掘(WUM)[2]。WUM 的研究内容是对 Web 日志数据进行分析，以发现潜在的有价值的访问模式。文献[3]为了降低数据规模,并从行为日志中发现更有推荐价值的访问模式,提出了基于用户兴趣特征的数据预处理方法。根据相关的研究，用户的短期行为是指用户并不熟悉其所访问的信息，由于对内容的不熟知性必然造成访问的随机性和无规律性。这些由于短期行为而产生的 Web 日志记录，其挖掘和推荐的价值是很小的[4]。从长期来看，网络用户所访问的网页往往落在相关的几个目录和网页集合上。为了验证网络用户的浏览分布规律，文献[5]分别对两组日志进行了分析，一组为 server log(1995 年 8 月的 NASA server log)，另一组为 Proxy log(国内某大学 2005 年 1 月 proxy log)，结果显示用户访问行为具有明显的重尾分布特征，必须去除一些噪音数据。

为了能够在不丢失有价值日志记录的前提下，有效降低 WUM 日志数据的规模，减小后期挖掘阶段的复杂度，文中将知识的信息量引入到数据预处理阶段，在日志记录形式化后，将每个不同 URL 看作属性，通过计算每个 URL 属性相对于属性集的重要性量化值，结合页面兴趣度，并采用操作系统中 LRU 页面置换算

① 基金项目:安徽科技学院青年基金(ZIC2011117);安徽科技学院教研课题(X201014)
收稿时间:2010-09-09;收到修改稿时间:2010-12-30

法的思想来过滤掉因为用户偶尔行为而产生的日志记录，因为这些噪音数据一方面没有规律可循，另一方面会覆盖用户真正兴趣中规律性的行为。

2 基于属性重要性的WUM数据预处理

2.1 相关定义和性质

我们首先建立信息系统中知识与信息量的关系，将 Web 日志数据进行形式化表示，并通过知识的信息量对属性的重要性进行定义，来计算每个 URL 属性相对于属性集的重要性量化值。

定义 1. 信息系统可表示为四元组 S=(U, A, V, H)，其中 U 表示为对象的非空有限集；A 表示属性的非空有限集合；V 为 A 的值域，V=$\bigcup V_a$,其中 a∈A；H 表示为一个 U×A→V 的信息函数，它为每个对象的每个属性赋予一个信息值，即∀ a∈A, x∈U, H(x,a) ∈Va。

定义 2. 对于每个属性子集 P⊆A, 存在了一个二元不可区分的关系 DIS(P):DIS(P)={(x,y) ∈ U × U|∀ a∈A, H(x,a)=H(y,a)}。

定义 3. DIS(P)是论域 U 上的等价关系， U 中所有等价于 x 的元素所组成的子集 u 为 U 之元素 x 依 DIS(P)的等价类。

定义 4. 设 S=(U,A,V,f)为一信息系统，P⊆A，U/DIS(P)={X_1,X_2……X_n}.知识 P 的信息量定义为:

I(P)= $\sum_{i=1}^{n}\frac{|X_i|}{|U|}\left(1-\frac{|X_i|}{|U|}\right)$.其中|X|表示集合 X 的基数，$|X_i|/|U|$表示等价类 X_i 在 U 中的概率。

定理 1. 设 S=(U,A,V,f)为一信息系统，P⊆A，若 U/DIS(P)⊂U/DIS(A)，则 I(P)＜I(A)。

证明：令 U/DIS(P)={X_1,X_2……X_n}. U/DIS(A)={X_1,X_2……X_m}.

因为 U/DIS(P) ⊂U/DIS(A)，所以 n＜m. 因此存在{1,2,3……m}的一个划分 E={E_1,E_2……E_n}满足 $X_i=\bigcup Y_j$, j∈E_i, j=1,2,3……n 且存在某个 E_{io}∈E 满足 $|E_{io}|$＞1,因此$|X_i|=\sum$ ｜Y i｜,j∈E_i 且 i=1,2,3……n.

$\sum_{i=1}^{n}|X_i|_2=\sum_{i=1}^{n}\left(\sum_{j\in Ei}|Y_i|\right)^2>\sum_{j=1}^{m}|Y_j|^2$. 所以我们可得 $\sum_{i=1}^{n}\frac{|X_i|}{|U|}\left(1-\frac{|X_i|}{|U|}\right)<\sum_{j=1}^{m}\frac{|X_i|}{|U|}\left(1-\frac{|X_i|}{|U|}\right)$. 所以 I(P)＜I(A).

由定理 1 我们可以得到定理 2：设 S=(U,A,V,f)为一信息系统，那么 U/DIS(P)=U/DIS(A)的充要条件是 I(P)=I(A).

定义 5. 设 S=(U,A,V,f)为一信息系统，a∈A，属性 a 的重要性我们定义为：IMP(a)=I(a)。

定义 6. 设 S=(U,A,V,f)为一信息系统，a,b∈A，属性 a,b 相对于 A 的重要性量化值定义为：$IMP_A(a)=I(A)-I(a)$，$IMP_A(b)=I(A)-I(b)$，如果 $IMP_A(a)$＞$IMP_A(b)$，则属性 b 相对于 A 而言比属性 a 更重要。

上述定义 6 表明：上述定义 6 表明：属性 a、b 相对于属性集 A 哪个更重要，可以通过 A 去掉 a 后的信息量和 A 去掉 b 后的信息量的大小来度量。

定义 7. S=(U,A,V,f)为一信息系统，C⊆A，任何属性 a∈A-C 关于属性集 C 的重要性定义为：IMPc(a)=I(C∪{a})-I(C)。

2.2 基于属性重要性的 WUM 数据预处理

我们先利用常规的数据清理方式对数据集进行初步处理，然后有别于传统数据预处理的步骤，我们对初步处理后的数据集先用户识别，然后进行形式化表示， 我们以 S=(U,A,V,f)的形式来表示 web 日志数据，再转换得到用户访问情况表。User={Useri|i=1,2……n}为所有用户的集合，用来表示 U 论域；Page={Pagei|i=1,2……n }为所有各异页面的集合， 用来表示属性集 A； V 为用户访问页面的次数。那么 S=(U,A,V,f) → (User,Page,V_A,f) 。具体模型为：S=(U,A,V,f) ， U={Useri|i=1,2, … ,n} ， A={Pagei|i=1,2,…,n}，V 为 VA 的集合。

在形式化后的用户访问情况表中，我们是以页面访问次数作为表中元素值来考察，为了能更精确地删除噪音数据，我们再结合页面兴趣度来考虑，由相关文献可知，兴趣度与页面浏览次数、页面停留时间和接收字节数三者是线性相关的[6]，我们可以通过这三个参数利用文献[6]介绍的线性回归法求出页面的兴趣度 I。当页面相对于页面集的重要性量化值(由定义 5 可知)大于给定阈值 minIMP 且页面的兴趣度小于给定阈值 minINT 时，则与该页面相关的访问记录可能是噪音数据。因为阈值 minIMP 和 minINT 选取的好坏直接影响到数据清理的效果，所以下面我们来确定其取值。

页面平均重要性 AIMP

如何确定阈值 minIMP 是值得研究的问题，因为每个页面相对于页面集合的重要性一般是各

不相同的，根据用户访问的历史数据分析，重要的页面其相对于页面集合的重要性量化值基本都在平均水平之上，令页面平均重要性为 AIMP，AIMP=$\sum_{i=1}^{m} IMP_A(P_i)$ /m(其中 m 为不同 URL 的个数)。所以我们用 AIMP 作为阈值 minIMP。

页面平均兴趣度 AINT

根据日志记录的访问历史分析，用户感兴趣的页面的兴趣度普遍较高，他们在一段时间内所访问的网页相对都集中在几个目录和网页集合内，而其余网页的访问则相对随机，兴趣度相对较低。所以我们采用页面的平均兴趣度 AINT=$\sum_{i=1}^{n} INT_i$ /n (其中 n 为不同 URL 的个数)作为阈值 minINT。

某页面在满足 minIMP 和 minINT 两个阈值的条件下，为了进一步提高清理数据的准确度，我们采用类似于操作系统中 LRU 页面置换算法的思想，如果该页面最近很久未被访问，则认为与该页面相关的访问记录为噪音数据，予以淘汰。这样在上述三个条件的限制下，发现噪音数据更为精确。因为某个页面虽然满足 minIMP 和 minINT 两个阈值的条件，但却是最近被访问的，则我们认为它可能是用户最近才产生的兴趣，很可能在不久的将来还会被访问，如果把该访问记录删除，很可能影响用户的访问连续性和规律性，显然是不合适的。为此我们为所有日志记录中不同的页面设置一个访问时间域，用以记录这个页面自上次被访问以来所经历的时间 T，并设置一时间阈值 minTIM，如果该页面的 T>=minTIM，则说明它最近很久未被访问，是噪音数据，应该剔除。

2.3 处理过程描述

输入：web 日志记录集

输出：除去噪音数据后的 web 日志集

Step1: 将 web 日志集转化为(User,Page,VA,f)形式的用户访问情况表

先将用户识别后的所有用户以及所有不同的 URL 读入二维数组 visit 第一列和第一行中;

```
For i=1 to m // m 为用户个数
For j=1 to n // n 为 web 日志记录的个数
if visit[i][0]与第 j 条 web 日志的用户匹配 then
  { 将 visit 中第 i 个用户相应的 URL 访问次数加 1;}
  End if
  End For
End For
```

Step2: 剔除噪音数据

计算每个页面 Pi 相对于页面集合的重要性量化值并存入二维数组 Important;

计算每个页面 Pi 的兴趣度并存入二维数组 Interest;

计算每个页面 Pi 的访问时间域 T 并存入二维数组 Time;

```
sum = 0 ,eui = 0;
For i=0 to n   // n 为所有不同 URL 的个数
  sum = sum + Important[i];
End For
minIMP = sum/n;
For k1=0 to n
  eui = eui + Interest[1][k1];
End For
minINT = eui/n;
For j1=0 to number // number 为日志记录的数量
  For j2=0 to n // n 为不同 URL 的个数
    if 第 j1 条记录的 URL<>Interest[0][j2] then
      if Important[1][j2]>minIMP&&Interest[1][j2] < minINT then
        if Time[1][j2]>=minTIM then
          Delete 第 j1 条记录;
        End then
      End then
    break;
  End then
End For
End For
```

3 实例演示

给定的信息系统 S=(User,Page,VA,f)，我们将采集到的如表 1 所示 web 日志记录形式化为信息系统 S 的数据，得到如表 2 所示的用户访问情况表。为了演示方便，其中 U={U_1,U_2,U_3,U_4,U_5}，表示 5 个用户，A={P_1,P_2,P_3,P_4,P_5}，表示访问的网址，V 为某个用户在访问过程中访问该页面的次数。并假设阈值 minTIM 为 30min。

表 1　web 日志记录

NO	IP	Time	URL	Agent
1	211.70.51.11	10/oct/2010:13:04:40	B.html	Mozilla/4.0(win+xp)
2	211.70.51.11	10/oct/2010:13:04:45	E.html	Mozilla/4.0(win+xp)
3	211.70.51.11	10/oct/2010:13:10:47	F.html	Mozilla/4.0(win+xp)
4	211.70.51.11	10/oct/2010:13:11:50	B.html	Mozilla/4.0(win+xp)
5	211.70.51.12	10/oct/2010:14:00:20	A.html	Mozilla/4.0(win+xp)
6	211.70.51.12	10/oct/2010:14:01:20	D.html	Mozilla/4.0(win+xp)
7	211.70.51.12	10/oct/2010:14:01:33	D.html	Mozilla/4.0(win+xp)
8	211.70.51.13	10/oct/2010:14:03:50	A.html	Mozilla/4.0(win+xp)
9	211.70.51.13	10/oct/2010:14:03:51	A.html	Mozilla/4.0(win+xp)
10	211.70.51.13	10/oct/2010:14:04:11	B.html	Mozilla/4.0(win+xp)
11	211.70.51.13	10/oct/2010:14:04:25	B.html	Mozilla/4.0(win+xp)
12	211.70.51.13	10/oct/2010:14:04:30	D.html	Mozilla/4.0(win+xp)
13	211.70.51.13	10/oct/2010:14:04:32	D.html	Mozilla/4.0(win+xp)
14	211.70.51.13	10/oct/2010:14:04:40	D.html	Mozilla/4.0(win+xp)
15	191.10.1.112	10/oct/2010:14:05:01	A.html	Mozilla/4.0(win7)
16	191.10.1.112	10/oct/2010:14:05:12	B.html	Mozilla/4.0(win7)

表 2　用户访问情况表

	P_1	P_2	P_3	P_4	P_5
U_1	0	2	0	1	1
U_2	1	0	0	2	0
U_3	2	2	0	3	0
U_4	1	1	1	1	1
U_5	1	0	0	2	0

(1) 首先我们根据文献 5 提出的算法求出各页面的兴趣度：INT_{P1}=69.58,INT_{P2}=59.27,INT_{P3}=12.26, INT_{P4}=83.12,INT_{P5}=16.75。

(2) 由定义 2 和定义 3 我们可以得到：U/DIS(A)={U_1,{U_2,U_5},U_3,U_4}，所以由定义 4 可知：I(A)=18/25。

(3) 根据定义 6 计算各属性(页面)相对于属性集合(页面集合)的重要性量化值：

$IMP_A(P_1)$=I(A)-I(P_1)=18/25-14/25=4/25.

$IMP_A(P_2)$=2/25.

$IMP_A(P_3)$=10/25.

$IMP_A(P_4)$=2/25.

$IMP_A(P_5)$=6/25.

(4) 根据上面提出的方法求出：

AIMP=28/125. AINT=48.19

(5)页面 P3 和 P5 的页面重要性量化值分别为 10/25 和 6/25，两者均大于 minIMP。而 P3 和 P5 的页面兴趣度分别为 12.26 和 16.75，两者均小于 minINT，如果 P3 和 P5 的时间访问域 T 都不小于 minTIM，则由上述可知，P3 和 P5 对应的访问记录为噪音数据，应删除。

4　实验结果

我们在 CPU 为 Intel(R)T2130 1.86GHZ，内存为 2G 和 Windows XP 对文章提出的方法进行了实验并与文献[3]的方法进行了比较，数据采用校园网的 Web 日志数据集。先使用通用的数据清理方法清洗日志记录得到数据集 a，然后再用本文介绍的方法进一步对数据集 a 进行过滤处理。

图 1 给出了提取某一比例(如：90%)的 web 日志记录，本文提出的方法与文献[3]方法在滤掉的噪音数据比例上的比较。我们可以看到本文的方式能更有效地发现噪音数据。但由图 2 可知：本文的清理方式所需时间要比文献[3]的方法多，随着提取日志比例(≤60%)的下降，两者时间上的差值会有所减小。图 3 则给出了分别将阈值 minTIM 设为 10～60min，减少的数据量的比例。

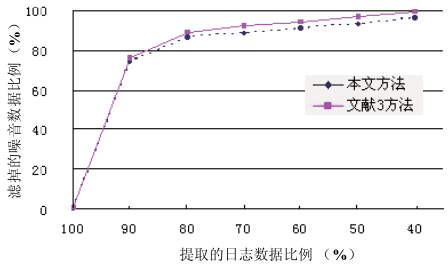

图 1　两种方法过滤掉的噪音数据比较

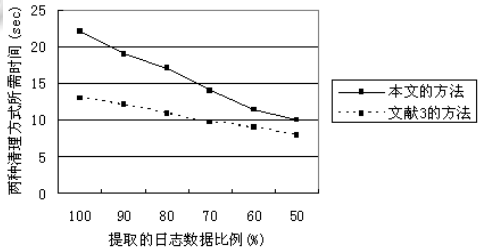

图 2　两种方法耗费的时间比较

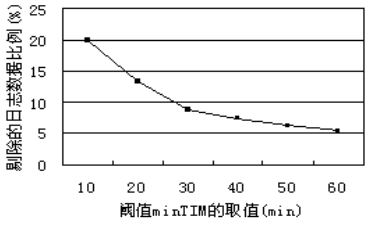

图 3　阈值 minTIM 的取值与过滤掉的噪音数据

（下转第 247 页）

深度搜索下一次扩展的是本次扩展出来的子节点中的一个，而广度搜索扩展的则是本次扩展的节点的兄弟节点。

广度优先搜索一般只用于找最优解，不会用于找所有解；深度优先搜索可以搜索出所有的解。

4.2 回溯法

回溯算法是所有搜索算法中最为基本的一种算法，其采用了一种“走不通就掉头”思想作为其控制结构，其相当于采用了先根遍历的方法来构造解答树，可用于找解或所有解以及最优解。

4.3 双向广度优先搜索

广度搜索虽然可以得到最优解，但是其空间消耗增长太快。但如果从正反两个方向进行广度搜索，理想情况下可以减少二分之一的搜索量，从而提高搜索速度。

4.4 分支限界

对于每个扩展出来的节点给出一个预期值，如果这个预期值不如当前已经搜索出来的结果好的话，则将这个节点(包括其子节点)从解答树中删去，从而达到加快搜索速度的目的。

4.5 算法

A*算法中更一般引入了一个估价函数 f，其定义为 f=g+h。其中 g 为到达当前节点的耗费，而 h 表示对从当前节点到达目标节点的耗费的估计。其必须满足两个条件：

1) h 必须小于等于实际的从当前节点到达目标节点的最小耗费 h*(n)。

2) f 必须保持单调递增。

A*算法的控制结构与广度搜索的十分类似，只是每次扩展的都是当前待扩展节点中 f 值最小的一个，如果扩展出来的节点与已扩展的节点重复，则删去这个节点。如果与待扩展节点重复，如果这个节点的估价函数值较小，则用其代替原待扩展节点。

参考文献

1 Kreher DL, Stinson DR. Combinationial Algorithms—Generation. Enumeration and Search. London: CRC Press, 1999: 151－186.

2 Jungnickel D. Graphs Networks and Algorithms. Translated from German by Tilla Schade Springer, 1999: 3－51.

3 吴文虎,王建德.实用算法的分析与程序设计.北京:电子工业出版社,1998.113－200.

4 严蔚敏,吴伟民.数据结构.北京:清华大学出版社,1999.

5 王晓东.算法设计与分析(C 语言版).北京:电子工业出版社,2001.162－191.

6 张德富.算法设计与分析.北京:国防工业出版社,2009.113－120,18－222,41－243.

（上接第 222 页）

5 结语

为了能有效降低后期挖掘的数据规模，尽量在不丢失重要信息的前提下降低发现访问模式的复杂度，本文提出了基于属性重要性的 WUM 数据预处理方式。该方法在引入知识信息量的基础上，给出了属性相对于属性集的重要性量化值的概念，并与页面兴趣度相结合，采用了 LRU 页面置换算法的思想，能有效地剔除噪音数据，降低 web 日志数据的规模。下一步笔者将进一步研究如何更好地设置三个阈值 minIMP、minINT 和 minTIM，并设法改进本文提出的方式，使其在时间效率上有更好的表现。

本文作者创新点：本文在引入知识的信息量的基础上，给出了单个属性相对于属性集的重要性量化值的概念，并采用了操作系统中 LRU 页面置换算法的思想，提出了基于属性重要性的 WUM 数据预处理方式。

参考文献

1 Tan PN, Steinbach M, Kumar V.数据挖掘导论.北京:人民邮电出版社,2006.

2 Mobasher B. Web usagemining and personalization. Chapman Hall & CRC Press, Baton Rouge, 2003.

3 杨明花,古志民.基于兴趣特征的 WUM 数据预处理方法.计算机应用,2006,26(10):133-134.

4 王春霞,王迤冉.WEB 日志挖掘实现网站优化.微计算机信息,2006(33):75-77.

5 陆丽娜,杨怡玲,管旭东,等.Web 日志挖掘中的数据预处理的研究.计算机工程,2000,26(4):66－67.

6 任永功,付玉,张亮.基于 web 日志的连续频繁路径挖掘算法.小型微型计算机系统,2008,29(12):2272－2276.