# 基于 Web 数据挖掘的一种个性化方法

石 军[1] 王儒敬[1] 王志红[2]
[1](中国科学院智能机械研究所,合肥 230031)
[2](华东电子工程研究所,合肥 230031)
E-mail:dmresearch@gmail.com

**摘 要** 文章应用 Web 数据挖掘的相关知识,对网站内容现有的资源内容设立相关度,根据网站用户日志,使用在线分析的方法对用户行为模式进行分析挖掘,根据网站类别内容的相关度预测用户的兴趣,为用户提供最合适更具人性化的信息资源。

**关键词** 个性化 Web 数据挖掘 信息过滤 用户模式

**文章编号** 1002-8331-(2006)07-0137-03 **文献标识码** A **中图分类号** TP391;TP393

# An Approach Web Mining Based Personalized Web

**Shi Jun[1] Wang Rujing[1] Wang Zhihong[2]**
[1](Institute of Intelligent Machines of Chinese Academy of Science, Hefei 230031)
[2](East China Institute of Electronic Engineering, Hefei 230031)

**Abstract**: In this paper, we set lots of similarity between every two classes of the website, then use Web mining and related technology, analyze the Web usage log and gain the user usage pattern, predict the user interest according to the classes' similarity, offer a more human webpage for each user.

**Keywords**: personalization, Web mining, information filter, user pattern

## 1 引言

个性化 Web 站点就是根据用户浏览模式和个人兴趣爱好以及用户资料,结合网站结构和内容专门提供相关信息的网站[1,2]。简单来说就是不明确咨询用户而利用其他资料提供用户感兴趣的信息。在 Internet 日益发展的今天,信息过载而知识贫乏是一个普遍存在的问题。根据我们在今年三月份做的一个月统计数据中得知:全球互联网搜索巨头 Google 收录国内几大门户站点的页面数目大致如下:

表 1 搜索引擎 Google 收录统计

| sina.com.cn | 163.com | sohu.com | qq.com | yahoo.com | msn.com |
|---|---|---|---|---|---|
| 5 930 000 | 2 980 000 | 2 720 000 | 367 000 | 23 100 000 | 7 220 000 |

在动辄数以百万计的网站中每个人想要找到自己感兴趣的资料内容是一件很困难的事情。传统的方法是使用站点搜索或者专门的搜索引擎以及超链来获取自己需要的知识,然而这些方法都存在一定的弊端,如使用搜索方法得到知识。当前的搜索方法都是基于关键字的,也就是说你在 sina.com.cn 站点找你所感兴趣的"数据挖掘"方向的资料,那么你得到的结果要么是包含有"数据挖掘"关键词的页面,要么是网站管理员根据自我定义将某些页面定义为数据挖掘方向的,实际页面的内容并不包含该关键词。由于网站管理员知识的欠缺,那么"聚类"、"分类"和"关联规则"等内容的资料可能你就不会发现;若你根据超链的方式冲浪,也很有可能存在类似的问题。

数据挖掘就是从大量的数据中抽取未知的、感兴趣的模式或规律等知识的方法[3]。数据挖掘技术在一定的程度上可以解决"信息过量而知识贫乏"的问题。Web 数据挖掘则是使用数据挖掘技术从 Web 文档中自动抽取和发现兴趣模式。Web 数据挖掘和信息检索以及 Web 信息抽取有一定的差别[4]。Web 数据挖掘技术包括三个方面的内容[4,5]:

(1)Web 内容挖掘(Web Content Mining,WCM):对网站的文本,图像,音频文件等类型数据进行挖掘;

(2)Web 结构挖掘(Web Structure Mining,WSM):主要研究内容在于站点和网页的链接信息,结合网站内容的组织结构的分析以及 Internet 内网站的超链分析;

(3)Web 使用挖掘(Web Usage Mining,WUM):就是利用数据挖掘技术对网站的用户访问数据(服务器日志)及其他相关数据的分析挖掘,并从中获取有价值的模式知识。

从数据挖掘算法方面考虑,使用分类和聚类算法对网站网页或者用户浏览的网页以及用户进行分类和聚类;在考虑网页之间相关性时可以使用关联规则,对于置信度和支持度较高的网页集还可以考虑在网站重构时对原本没有建立的超链页面新建立超链接以方便用户访问;对于用户访问网站的页面序列可以使用时间序列模式来分析预测用户的访问趋势。

在本论文中我们使用 Web 内容挖掘技术和 Web 使用挖掘技术,并且利用了数据挖掘中的关联规则及其他相关算法进行网站的个性化设计。该模型应用于数据挖掘研究院网站,根据网站对用户进行的相关调查结果显示基本能够满足访问用户的需求。本论文组织结构分别为:第 2 节对比当前网站个性化所使用的技术方法;利用数据挖掘研究院(China Data

**基金项目**:国家 863 高技术研究发展计划资助项目(编号:2001AA118070)

**作者简介**:石军(1978-),男,硕士研究生,主要研究方向:Web 数据挖掘,Web 信息抽取。王儒敬(1964-),男,研究员,主要研究方向:智能决策支持系统。王志红(1979-),女,硕士研究生,研究方向:数据处理。

Mining Research) 资源设计的基于 Web 数据挖掘技术个性化网站系统 DMRP(Data Mining Research Personalization)在第 3 节进行描述;第 4 节对我们所建立的 DMRP 系统进行分析;第 5 节提出对今后工作的展望。

## 2 相关研究

人们对互联网中应用个性化技术有一定的研究,也取得了一些成果。曾经在 CMU 大学计算机主页上服役过的 WebWatcher 根据用户访问模式使用了基于 agent 的智能引导技术来帮助用户更好地访问[5,6];在文献[7~10]中使用了一种称为协同过滤(Collaborative Filtering)的方法,它的基本假设是经常访问相似资源的用户兴趣相似,相似兴趣的用户又会访问相似的资源。因此,通过对相似兴趣用户的判定,来确定某个用户对某一未知资源是否感兴趣;另外一种重要的方法就是基于内容的过滤方法(Content-Based Filtering),它是根据用户访问网站的历史模式分析挖掘对用户兴趣的不断学习和反馈,以保证过滤后的信息内容和用户的兴趣相吻合,这在 NewsWeeder[12], Ringo[13], Lira[14], IDD News Browser[15], PRES[16]等系统都有研究分析;Fab[17], TREC-8[18]等则是使用协同过滤和基于内容的过滤两种方法结合实现网站个性化,也就是二者的混合方法;个性化网站技术方法还有在文献[1]中提出的基于规则的过滤,就是网站对用户进行一系列的提问,问题以决策树的形式存在,用户逐一回答,最后将会返回给用户一个比较适合的答案。

基于假设的协同过滤方法考虑整体性,一些研究人员提出了改进方法如在文献[19]中使用协相关系数,基于向量的相似度计算以及统计贝叶斯方法;在文献[20]中提出基于页面项集的改进方法,但是对于每个具体的用户考虑得不够,那样实现的个性化在某种程度上缺少人性化;内容过滤的方法则侧重于用户资料和用户历史访问信息,对于单个用户来说为了个人隐私问题有一定的欺骗性;基于规则的过滤需要用户配合的过多,很少用户会完整回答问题;当前较好的方法是结合协同过滤与基于内容的过滤的混合方法来实现网站的个性化设计。

也有一些研究人员从其他方面对个性化网站的实现做了大量研究,其中包括 Web 数据挖掘和个性化网站实现的数据处理问题[21~23];Honghua Dai 等[24]使用了将域本体和 Web 使用挖掘以及个性化相结合的方法。Corin R.Anderson 等提出了面向手机、PDA 等无线接入设备的网站个性化方法[25]。

我们使用了混合信息过滤方式,根据人工设立初始类别相关度,建立了一套可行的系统模型来实现网站个性化。这种方法和以前的研究相比,我们的系统框架简化了一些计算,将离线操作与实时推荐相结合,在不增加现有网站硬件设备的同时使网站系统更具人性化。

## 3 DMRP 系统

### 3.1 系统架构

DMRP 系统是一种在线分析用户日志并实时推荐的系统。该系统设计的相关结构图如图 1。

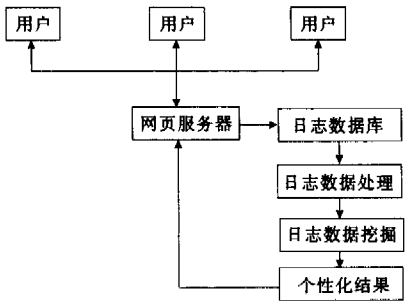


图 1 DMRP 系统设计框架

### 3.2 日志数据处理

当用户请求 Web 页面时,Web 服务器启动日志记录器记录用户日志,对该用户的日志进行过滤处理,如判断该用户为正常访问用户还是 Crawler 程序。随后对正常访问用户进行历史浏览行为模式的分析得到该用户的兴趣内容集,随后生成页面返回给用户;若为 Crawler,那么它的访问模式必定是一种异常模式,由于 Crawler 程序对网站超链的访问不像人那样具有选择性,他都是对系统的所有类别标签都会进行访问,而一般用户则只是访问自己感兴趣的类别,所以我们定义的异常模式为当某个客户端访问的类别数目占系统类别数目的 10%时就认为该客户端为 Crawler 程序。

由于对于用户回话的辨识一直都没有一种很好的解决办法。当前可使用的方法有:(1)在客户端建立 Cookie;(2)使用用户回话 ID;(3)使用用户访问 IP。但是这些方法都有弊端。由于操作系统安全机制可以屏蔽客户端 Cookie 的建立,另外用户可以删除客户端建立的 Cookie;若使用用户回话 ID 作为辨识标志的也不准确,可能客户端安装了不同的浏览器,若几个浏览器同时使用那么将会在系统形成不同的回话 ID,而物理上仍然是一个用户访问系统;大部分用户都是由同一个网关进入互联网,所以根据 IP 来辨识也是不科学的办法。鉴于这些辨识方法的特点,我们在 DMRP 系统中同时使用了三种用户会话辨识方式。首先在用户客户端建立 Cookie,如果客户端不允许建立 Cookie 或者建立的 Cookie 被用户删除,那么将使用用户回话 ID 和用户访问 IP 相结合的办法。在此可能很多用户是使用同一个 IP 访问,我们认为同一个 IP 下的用户属于同一个兴趣组,也就是使用同一个公共 IP 访问 Internet 的用户他们的兴趣大部分是相同的,在现实中使用同一个 IP 上网的用户他们可能是同一个研究实验室,同一个学校或者同一个社区,因此兴趣相同的概率也较大。我们在客户端建立的 Cookie 并不是用户在系统中的帐号,而是访问系统时系统分配给他的一个识别码,这是因为考虑到可能有的用户并没有在网站注册和注册了的用户访问网站不一定必须要登录才能看到内容。

### 3.3 模型算法

日志其他记录内容按照 W3C 的标准建立,还有上页 URL,本页 URL,回话 ID,用户 IP,访问页面时间等记录项。在此将日志记录项记作:

item{ID,Referrer,NowURL,SessionID,ClientIP,Visit_time}

数据库中日志数据对应也就是:

Log{item1,item2,…,item$N$}

我们在每建立一个页面时都对该页面按照内容分类并贴上系统标签加以标记。系统页面标签集记作:

Label{label1,label2,…,label$M$}

我们对标签之间的相关性默认一个相关度,整个标签集构成一个稀疏的有向图。如 label1="聚类",label2="分类",label3="数据挖掘论文",label4="数据挖掘新闻",初始我们凭借经验知识建立如图 2 所示关系:

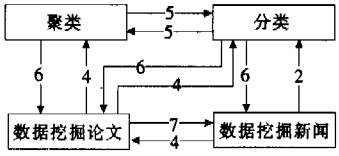


图 2 类别相关示例

在这里我们定义最大相关度区间为[0,1),因为我们认为完全相关的不会存在,所以在系统中并不存在相关度为 1 的两个类别。之所以构成系数图是为了减少系统的复杂性,因为对于一些我们认为几乎不存在相关问题的类别可以不建立相关度,也就可以认为相关度为 0。建立有向图是根据经验知识而设,对于访问数据挖掘论文的用户(对其感兴趣)访问不会一定访问分类或者聚类,但是由于对聚类感兴趣的用户由于聚类本身的特点,用户对数据挖掘论文感兴趣的可能性比较大。因此在这里定义:

Sim{label1,label3}=0.6,Sim{label3,label1}=0.4

**定义 1** 用户对 *Label*[*i*]类别的主兴趣度就是用户对 *Label*[*i*]类别下内容访问的次数。主兴趣度 *L*[*i*]对应公式为:

$L[i]=Label[i]\times C[i]$(在这里 $C[i]$为类别 *Label*[*i*]用户点击数)

**定义 2** 用户对 *Label*[*i*]的附加兴趣度就是用户访问所有其他类别的主兴趣度与 *Label*[*i*]的相关度的乘积乘以一个衰减因子。其公式为:

$J[i]=c\times J[i]\times s[j,i]$($c$ 为衰减因子)

在这里,$J[i]$为用户对 $i$ 类别的主兴趣度,$s[j,i]$为 $j$ 类别与 $i$ 类别的相关度;很明显用户对 *Label*[*i*]类别访问的次数越多那么说明该用户对用户的兴趣度(主兴趣度)就越大;根据定义 2,用户对 *Label*[*i*]的兴趣度越大,那么在类别标签集中与 *Label*[*i*]相关度越大的类别用户感兴趣的可能就越大,在这里 $J[i]$也就越大。根据用户历史访问模式,用户对网站标签 *Label*[*i*]的兴趣度 $Z[i]$为:

$Z[i]=I[i]+J[i]$

随后对用户的兴趣度集从大到小排序,从中选择 $n$ 个类别(在这里我们系统选择 $n$=4),再在每个类别中选择用户未曾访问的一定数目的内容(系统定义每个类别取 3 条)生成新的基于用户兴趣的个性化 Web 网页内容。

### 3.4 参数修正

由于标签的相关度是我们根据经验知识定义的,肯定在一定程度上存在偏差和失误。在此我们对推荐给用户的个性化页面记录用户访问模式,若用户对推荐的内容感兴趣,那么对于用户访问推荐内容做记录, 并对该类别的该内容做一定的补偿;若用户对推荐页面的内容都不感兴趣,那么就对所有推荐类别的内容进行惩罚。根据所有用户访问的模式我们定期对补偿和惩罚的数据进行处理,用这些数据来修正我们开始默认的类别之间的相关度。

## 4 DMRP 系统分析

DMRP 是一种基于 Web 数据挖掘的使用了混合过滤的方法建立的系统模型。我们使用了关联规则和人工定义的方法计算网站内容的相关度;使用基于内容的过滤方法计算用户对网站内容的兴趣度顺序;协调过滤的方法则用在修正相关度方面。

根据第 3 节算法和 DMRP 系统相关内容统计数据如表 2。

由表 2 中的内容我们可以知道 DMRP 具有很好的扩展性,对于大型的门户网站只需要使用,使用关联规则和聚类、分类等联合方法确定类别之间相关度的问题;由于大型网站多采用分布式多主机的方式,因此单个服务器的压力并不比我们测试服务器大,因此在数据库速度上也是可行的。对于使用协同过修正类别相关度可以采用离线方式处理。

表 2 DMRP 系统统计数据

| 名称 | 内容 |
|---|---|
| DMRP 数据库操作 | |
| 用户访问一次页面记录日志数 | 2 条 |
| 有效日志数据处理生成数据 | 1 条 |
| 生成个性化页面对数据库查询次数 | $2*n^2$($n$ 为用户访问网站类别标签数) |
| DMRP 网站数据 | |
| 日平均访问用户 | 2 000(独立 IP) |
| 日平均页面浏览次数 | 50 000 |
| 日平均生成个性化页面次数 | 10 000 |
| 日最高峰在线日数 | 200 |
| 网站标签数目 | 92 |
| 网站内容数目 | 850 |
| 用户对 DMRP 系统意见调查 | |
| 基本满意 | 57% |
| 很满意 | 2% |
| 效果很差 | 0% |
| 其它(包括对网站资料数目少不满意等) | 41% |

## 5 今后工作展望

在人们日益感觉到被信息淹没而知识贫乏的时候,网站个性化设计有助于用户的信息获取以及提高网站用户的忠贞度。本文提出了一种简易实用的技术模型,有利于推动网站在线实现网站的个性化。由于网站信息量一直都以几何级数增长,因此对于实现在线数据挖掘的个性化网站技术性能需要继续提高;在今后我们会研究本体与数据挖掘技术相结合计算类别相关度的方法,如相关度的传递问题;对于网页内容的分类将会研究使用本体与聚类、分类知识相结合的多标签标记方法;当前 tag 日益在各大网站盛行,我们还会尝试利用一些标准(如 W3C 的标准日志)开发开放源代码的软件包推动个性化互联网的发展。(收稿日期:2005 年 10 月)

## 参考文献

1.Eirinaki M,Vazirgiannis M.Web Mining for Web Personalization[J].ACM Transactions on Internet Technology,2003;3(1):1~27

2.http://www.personalization.org/faqs1.html

3.J Han,M Kamber.Data Mining Concepts and Techniques[M].Beijing:High Education Press,2001

4.Kosala R,Blockeel H.Web mining research:A survey[J].ACM SIGKDD Explorations,2000;2(1):1~15

5.S K Madria,S S Bhowmick,W K Ng et al.Research issues in web data mining[C].In:Proceedings of Data Warehousing and Knowledge Discovery,First International Conference,DaWaK'99,1999:303~312

6.Mladenic D.Personal WebWatcher:Implementation and Design[R].Technical Report IJS-DP-7472,1996

7.T Joachims,D Freitag,T Mitchell.Webwatcher:A tour guide for the world wide web[C].In:International Joint Conference on Artificial Intelligence,1999-08

8.Konstan J,Miller B,Maltz D et al.GroupLens:applying collaborative filtering to use net news[J].Communications of the ACM,1997;40(3):77~78

(下转 168 页)

(12)else

(13) 输出:无变化发生;//即当 $P(C_i|X)\geqslant\alpha$ 时

(14)end;

(15)重复执行该循环指定的次数。

算法的空间复杂度即为 NBCC 的规模。

算法的时间复杂度分析如下:

**引理** 假设训练样本集分为 $m$ 类,样本有 $n$ 个属性,并且循环重复执行了 $c$ 次,则求解算法 NBCC 的时间复杂度为 $O(mnc)$。

## 4 性能分析

### 4.1 实验环境

实验环境:OS 是 Windows 2000,CPU 为 P Ⅳ,主频为 2.4GHz,主存为 512MB。算法实现的工具是 VC。

实验数据:IBM 的数据产生器 #assocSynData 产生的模拟数据集。

### 4.2 实验结果

表 1 取 $\alpha$=0.4

| 样本 | 类别 | | | |
|---|---|---|---|---|
| | $C_1$ | $C_2$ | $C_3$ | 新类 |
| $X_1$ | 0.56 | 0.96 | 0.43 | NO |
| $X_2$ | 0.89 | 0.63 | 0.36 | YES |
| $X_3$ | 0.79 | 0.52 | 0.21 | YES |

表 2 取 $\alpha$=0.3

| 样本 | 类别 | | | |
|---|---|---|---|---|
| | $C_1$ | $C_2$ | $C_3$ | 新类 |
| $X_4$ | 0.91 | 0.37 | 0.75 | NO |
| $X_5$ | 0.32 | 0.63 | 0.89 | NO |
| $X_6$ | 0.79 | 0.19 | 0.41 | YES |

说明:在上面的两个表格中,将 $t$ 时刻利用精确抽样得到的概要数据结构看作样本点 $X_t$ 的集合,其中,$t$ 为有限的自然数。

表 1 和表 2 表明:算法 NBCC 是有效和可行的,但是对阈值 $\alpha$ 的选取是敏感的。

## 5 结论及展望

本文讨论了在数据流环境下进行知识类型"变化"的挖掘过程,提出了挖掘"变化"的相应算法。结果证明该方法是可行的。

针对数据流环境中"变化"的分析和挖掘是区别于传统的基于静态数据集的分析和挖掘的显著特征,这个问题已经引起了国内外许多学者的关注,例如,被提出的新的并且有意义的研究题目有:概念转移研究、聚类变化研究等。

(收稿日期:2005 年 6 月)

## 参考文献


1.Brian Babcock,Shivnath Babu,Mayur Datar et al.Models and issues in data stream system[C].In:PODS'02,Madison,WI,2002-06

2.M Garofalakis,J Gehrke,R Raastogi.Querying and mining data streams: You only get one look[C].In:VLDB'02,HongKong,China,2002-08

3.Guozhu Dong,Jiawei Han,Laks V S Lakshmanan et al.Online mining of changes from data streams:research problems and preliminary results[C].In:ACM SIGMOD MPDS'03 San Diego,CA,USA,2003

4.Yixin Chen,Guozhu Dong,Jiawei Han et al.Online Analytical Processing Stream Data:Is It Feasible? DMKD 2002

5.Vitter JS.Random sampling with a reservoir[J].ACM Trans on Mathematical Software,1985;11(1):37~57

6.Gibbons PB,Matias Y.New sampling-based summary statistics for improving approximate query answers[C].In:Haas LM,Tiwary A eds. SIGMOD 1998,Proc of the ACM SIGMOD Int'l Conf on Management of Data,Seattle:ACM Press,1998:331~342

7.Jiawei Han,Michaeline Kamber 著.范明,孟小峰等译.数据挖掘概念与技术[M].北京:机械工业出版社,2001:196~199

8.盛骤,谢式千,潘承毅.概率论与数理统计[M].第 2 版,北京:高等教育出版社,1989:23~25


(上接 139 页)


9.Bamshad Mobasher1,Honghua Dai,Tao Luo et al.Combining Web Usage and Content Mining for More Effective Personalization[C].In: Proceedings of the International Conference on E-Commerce and Web Technologies(ECWeb2000)

10.J Herlocker,J Konstan,A Borchers et al.An algorithmic framework for performing collaborative filtering[C].In:Proceedings of the 1999 Conference on Research and Development in Information Retrieval, 1999-08

11.P S Yu.Data mining and personalization technologies[C].In:Int'l Conference on Database Systems for Advanced Applications(DASFAA99), Hsinchu,Taiwan,1999-04

12.Lang K.Newsweeder:Learning to Filter Netnews[C].In:Proceedings of the 12th Int Conference on Machine Learning,Stanford,California, 1995:331~339

13.Shardanand U,Maes P.Social information filtering:algorithms for automating word of mouth[C].In:Proceedings of the ACM CHI Conference,1995

14.M Pazzani,L Nguyen,S Mantik.Learning from hotlists and coldlists: Towards a WWW information filtering and seeking agent[C].In:Proc AI Tools Conf,Washington,DC,1995

15.Eric Bloedorn,Inderjeet Mani,T Richard MacMillan.Representational Issues in Machine Learning of User Profiles[C].In:Proceedings of the 1996 AAAI Spring Symposium on Machine Learning in Information Access(MLIA'96),Stanford,CA,AAAI Press,1996

16.Meteren R V,Someren M V.Using Content-Based Filtering for Recommendation[C].In:Proceedings of ECML Workshop:Machine Learning in New Information,2000:47~56

17.M Balabanovic,Y Shoham.Fab:Content-based,collaborative recommendation[J].Communications of the ACM,1997;40(3):66~72

18.Soboroff I M,Nicholas C K.Related,but not Relevant:Content-Based Collaborative Filtering in TREC-8[J].Information Retrieval, 2002;5(2-3):189~208

19.Breese J S,Heckerman D,Kadie C.Empirical Analysis of Predictive Algorithms for Collaborative Filtering[C].In:Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence,1998:43~52

20.Item-based Collaborative Filtering Recommendation Algorithms

21.Robert C,Bamshad M,Jaideep S.Data Preparation for Mining World Wide Web Browsing Patterns

22.Miriam Baglioni,U Ferrara,Andrea Romei et al.Preprocessing and Mining Web Log Data for Web Personalization.AI*IA,2003:237~249

23.Doru Tanasa,Brigitte Trousse.Advanced Data Preprocessing for Intersites Web Usage Mining[J].IEEE Intelligent Systems,2004;19(2): 59~65

24.Honghua(Kathy)Dai,Bamshad Mobasher.A Road map to More Effective Web Personalization:Integrating Domain Knowledge with Web Usage Mining

25.Corin R Anderson,Pedro Domingos,Daniel S Weld.Personalizing Web Sites for Mobile Users[C].In:Proceedings of the 10th Conference on the World Wide Web(WWW10),2001