# 个性化推荐推荐系统中基于 WEB 的挖掘①

汪彦红，杨波，胡玉鹏

(湖南大学 信息科学与工程学院，长沙 410082)

**摘 要**：Internet 的普及和应用带来了 WEB 上的信息爆炸，如何基于 WEB 挖掘技术设计有效的信息推荐算法和推荐系统成为当前的研究热点。开发了一种基于 WEB 使用的推荐系统 WRS（Web Recommendation System），在该系统中，提出了一种利用图形分割技术聚类用户访问模式的算法，并采用最长公共子序列算法对用户目前的行为进行识别。理论分析和实验结果表明，改进后的模型在推荐质量上有了较大提高。

**关键词**：个性化推荐系统；WEB 使用挖掘；访问模式挖掘

## WEB-Based Mining in the Personalized Recommendation System

WANG Yan-Hong, YANG Bo, HU Yu-Peng

(College of Information Science and Engineering, Hunan University, Changsha 410082, China)

**Abstract:** Due to the rapid development and wide applications, Internet has led to the information explosion on the WEB. It becomes research hotspots at present how to design effective algorithms and systems based on the technology of WEB Usage Mining. In this paper, we develop a recommendation system called WRS (Web Recommendation System), which is based on the application of WEB. In WRS, we propose a novel algorithm that makes use of the technology of image segmentation to cluster access modules, and adopts the parallel longest common subsequence algorithm to discern users' behaviors. Theoretical analysis and laboratory result show that our system is more effective and the recommendation performance is improved after using the new method.

**Key words:** personalized recommendation system; WEB usage mining; navigation pattern mining

随着 WEB 技术的高速发展，电子商务得到了前所未有的发展，互联网上的海量信息对用户和网站管理者都带来巨大的挑战，为了迎合用户多方面的信息需求，随着信息量的增多，信息/产品过载现象日益严重，给客户和商家都带来了诸多不便[1]。对 WEB 访问模式的分析以及建模有助于帮助用户方便的浏览自己所需要的信息，目前在个性化推荐中已经成为一个热点[2]。

运用数据挖掘技术对 WEB 服务器上的日志数据和交易数据进行挖掘，从中抽取感兴趣的模式，能够更好地理解客户的访问行为、改进站点结构、为客户提供更多的个性服务，以得到较高的推荐精确度[3]。

Huan and Kamber[4]提出了 WEB 挖掘，WEB 数据挖掘就是利用数据挖掘技术，自动的从 WEB 文档以及服务中发现和抽取信息的过程。WEB 挖掘可分为三类[5]：WEB 内容挖掘、WEB 结构挖掘，WEB 使用挖掘。

在 WEB 使用挖掘中，基于 WEB 访问日志文件和用户会话的数据分析及用户访问模式的挖掘主要有以下几种方法[6]：(1)统计分析，(2)关联规则，(3)聚类，(4)分类，(5)序列模式，(6)建立模型。

为了有效的提供在线预测，本文主要做了如下工作：第一，本文开发了基于 WEB 使用和组合算法的推荐系统 WRS（Web Recommendation System）。第二，本文提出了一种新的用户访问模式识别算法，该算法基于图形分割技术在挖掘过程中建立用户访问模式模

① 基金项目:湖南省自然科学基金(10JJ4042)
收稿时间:2011-02-19;收到修改稿时间:2011-03-18

型，并使用最长公共子序列算法[7]识别用户当前会话，进而得到用户会话序列模式，从而对系统预测提供了必要的条件。

## 1　WRS系统设计

本文提出了一种新的推荐系统架构叫做 WRS，这个模型分成交叉存取两个部分：离线部分和在线部分。图 1 阐述了 WRS 的结构。

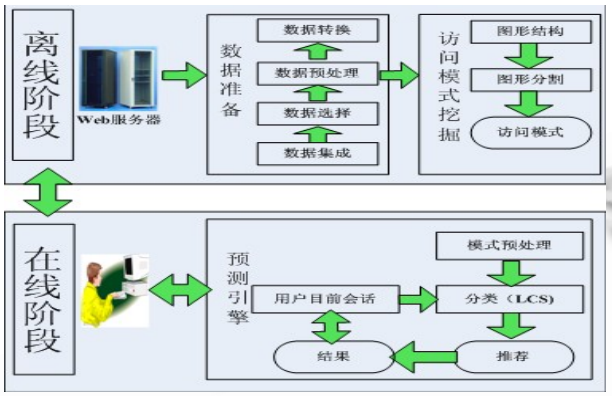


图 1　WRS 结构图

### 1.1 离线模块

离线模块的主要任务是通过挖掘引擎对用户事务数据库中的数据进行各种算法的挖掘，并产生相应的模式规则集，从中分析出有用的模式集提供给在线模块。数据预处理单元从用户原始的 WEB 日志文件中提取用户会话代码，并开发了一种建立在图形分割基础之上的新的聚类算法进行访问模式挖掘。

数据准备模块将采集模块收集到的数据进行处理，数据在预处理后将它们保存到用户事务数据库中，以便进一步挖掘。在数据分类之后，将数据文件转化为关系数据库，接下来的工作就是数据转换和聚合数据运算。

数据转换：将数据转换成一个分析模型。该分析模型是针对数据挖掘算法建立的。数据转换工作就是把一系列的网址以及关联信息转换成表格存储在相关数据库中，在此项工作中，网址数据库中关联了许多有价值的信息以减少实验复杂度，同时使用相关的数据库作为训练样本和测试样本来进行系统评估。

聚合数据运算：为用户会话和访问提供新的参数。这些数据在所应用的方法中具有统计学上的价值。例如，可以为整个数据库计算如下的参数：1、用户和会话的数量；2、在会话中重复访问网页的数量；3、会话长度所占的百分比。

在此系统中，将用户普遍的浏览特性描述成为用户访问模式。因为大量的用户在他们访问模式可能在某一个点有普遍的兴趣。访问模块应该去抓取重叠的爱好或者用户的需求信息。用户访问模式可以通过聚类算法被获得。聚类技术将数据对象按特征相近的原则划分为多个类或簇。在 WEB 使用挖掘领域有两种有趣的聚类：用户聚类和页面聚类[8]。当前的会话通过聚类算法就可以产生用户当前的访问模式，在线模块通过此模式进一步来分类用户目前的活动，用户访问模式被定义如下：

定义 1. 一个用户的访问模块 NP 记录了基于其感兴趣信息需求的整合的行为。会话聚类的结果 NP=(np1, np2,. . . , npk)用来描述用户一些的访问模式，每一个 npi 都是 p 的子集，p 则是网页的集合。

本文利用无向图分割算法实现聚类模型，以计算网页和所产生的邻接矩阵的相关度。它们分别是时间关联性和频率。时间关联性 TCa,b 衡量了一次会话对两个网页的访问需求度：

$$TC_{a,b}=\frac{\Sigma_{i=1}^{N}\frac{T_i}{T_{ab}}\times\frac{f_a(k)}{f_b(k)}}{\Sigma_{i=1}^{N}\frac{T_i}{T_{ab}}}$$

Ti 表示在第 i 次会话中访问 a 和 b 网页的总共时间，Tab 则是其时间差。频率 FCa,b 用来衡量两个网页在各自会话中的分布：

$$FC_{a,b}=\frac{N_{ab}}{\max\{N_a,N_b\}}$$

Nab 表示同时包含网页 a 和 b 的会话的个数，Na 和 Nb 分别表示仅仅包含网页 a 或 b 会话的数目。本文采用调和中项的方式表示在无向图中每一条边的权重。

$$W_{a,b}=\frac{2\times TC_{ab}\times FC_{ab}}{TC_{ab}+FC_{ab}}$$

本文将网页设置为顶点产生一个无向图，相当于邻接矩阵。顶点 a 和 b 之间的权值通过 Wa,b 来表示，为了限制边的数量，无向图边的权重必须大于阈值否则将被舍弃。本文中该阈值用 MinFreq 表示。

聚类算法的代码如下：

Input:

```
.Cleaned,filtered,and sessionized Log file.
.MinFreq.
.MinClusterSize.
Output:
. A list of clusters C
L[p]=p；//建立数据库联系
for each (pi,pj)L[p] do
  M(i,j)=WeightFormula(pi,Pj);
Edge (i,j)=M (i,j);
for    all Edge (u,v)Graph(E,V) do
//去除权值小于 MinFreq 的边
  if    Edge (u,v)<MinFreq then
remove (Edge (u,v));
for all vertices(u)Graph(E,V) do
        Cluster [i]=DFS(u); //使用 DFS 算法
If cluster[i]<MinClusterSize
//去除数量小于 MinClusterSize 的聚类
    Remove (Cluster[i]));
      end if
      i=i+1;
return (Cluster);
```

图像 2 阐述了聚类产生的例子。每一个网页在图中成为一个顶点，通过这些节点之间的相关度来建立一个无向图，因为有两个参数的限制，所以可以去掉一些边，然后形成几个聚类，再通过 DFS（深度优先搜索）算法在无向图中去发现相关部分。

设定两个参数分别为：MinFreq=0.2，最小聚类包含节点的数目为 3，剔除掉一些不符合条件的边，得到三个聚类。即访问模式挖掘最终所产生的结果如下：

$C_1 = NP_1 = \{P_1, P_2, P_5, P_6, P_9, P_{11}\}$

$C_2 = NP_2 = \{P_4, P_8, P_{14}\}$

$C_3 = NP_3 = \{P_7, P_{10}, P_{12}\}$

### 1.2 在线模块

通过系统在离线各个不同的阶段，访问模块已经出现，预测引擎中的在线模块包含着去预测用户未来需求的任务，在线模块最主要的就是预测引擎。

预测引擎被用来识别用户导航模块和预测推荐。为了识别用户当前的会话，本文需要寻找用户模式包含最大数量的相似网页在每一个会话中，LCS（最长公共子序列）算法用来去发现最长子序列。第二个目标为当前的会话计算推荐的内容，这些链接可能是基于相似用户惯用模式可能想去访问的。

使用 LCS 的推荐算法：

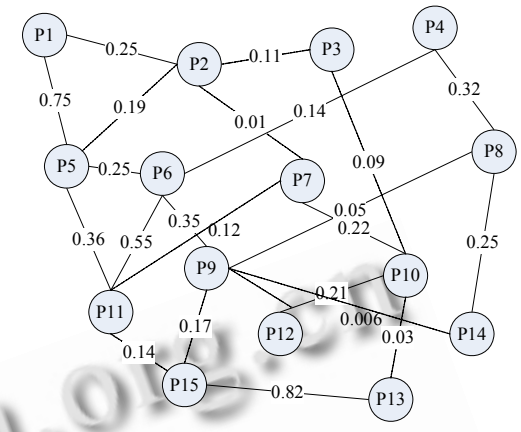


图 2 聚类的产生

（1）对推荐进行数据预处理：

在离线阶段对数据库中网站及其关联信息进行一系列的简化，进行数值定位简化，访问模块和聚类的划分，以及节点的表示和权重。

（2）基于 LCS 的用户识别：

输入有两个参数访问模式和有效对话窗口，访问模式通过 LCS 算法将发现最高相似度去预测用户下一步的行动并产生一个推荐序列。

（3）预测用户下一步的意向：

剔除不符合需求的推荐。将推荐序列推荐给用户。

推荐算法的代码如下：

```
Input:
. User’s active session window S.
. A set of navigation patterns np.
Output:
. A set of web pages as recommendation.
Sort (S); //  当前会话的识别分类
For each npinp do
   Sort (npi); //访问模式的识别分类
For each npinp do
AnswerSet=|LCS(npi,s)|;  //在目前会话和访问模式中得出 LCS
Mach_String=NP_String (Max (AnswerSet));
Prediction_Set=S-Mach_String;   //产生预测序列
Recommendation_set=Rank  (Prediction_Set);  //推荐给用户
Return (Recommendation_Set);
```

## 2 实验评估

多种技术被用来测量推荐系统的表现，相关实验许多都已经实现，WRS 通过在其在线阶段和离线阶段的参数来评估其效果。

### 2.1 实验评估标准

算法结果评价也有多个公式可以选择[9,10]，如 MAE、精确度、覆盖率等 。在个性化推荐系统目前的应用环境中，系统需要产生一个较小长度的电子商品推荐列表，而 MAE 等都是基于所有项目的用户偏好预测来评价算法的，不符合要求，选用准确率和召回率作为算法评价指标。准确率(Accuracy)表示既符合用户需求又被系统推荐的资源项的数目占系统推荐的资源数目的比例，覆盖率（Coverage）是指在推荐的内容中用户得到的推荐数和用户总的推荐请求数的比值，为了平衡两者，通常采用综合评价指标 F1 度量：

$$F1=\frac{2\times accuracy\times \mathrm{cov}\ erage}{accuracy+\mathrm{cov}\ erage}=\frac{2}{1/accuracy+1/\mathrm{cov}\ erage}$$

### 2.2 实验数据的选取

用户访问模式挖掘本文所选用的数据库是来自于自美国华盛顿大学提供的一个乐器资料的网站(http://machines.hyperreal.org)从 1997 年 2 月 12 日到 1999 年 4 月 30 号的服务器日志文件。日志文件格式为标准的服务器日志格式；另外一个数据库来自一家销售儿童书店的网站(CBW)。本文采用 Baraglia 和 Silvestri 所提出的聚类分析算法结果和 WRS 产生的结果进行了比较。本文用上面介绍到的两个参数来做如下实验：

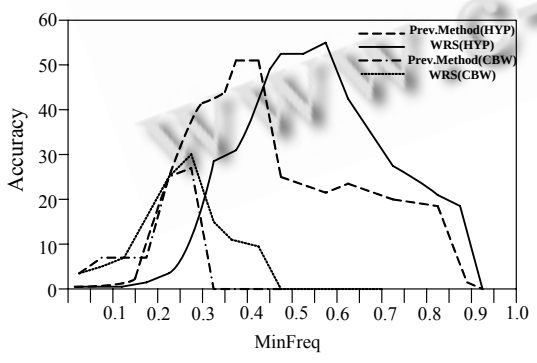


图 3　MinFreq 变化对精确度的影响

图 3 描述了 MinFreq 从 0 到 1 的变化过程当中系统在 HYP 和 CBW 数据库上所取得的推荐精确率，实验表明跟以前的方法相比，当有许多不同 MinFreq 阈值的时候此 WRS 系统可以有效的提高推荐的精确率。

由于精确率和覆盖率不能整体衡量推荐结果的适用性，所以本文采用了前面提高的综合参数 F1，通过图 4 可以看出 WRM 系统所产生 F1 参数在两个数据库上的表现均优于以前方法，其表明推荐结果比较合理，达到了 WRM 的目标。

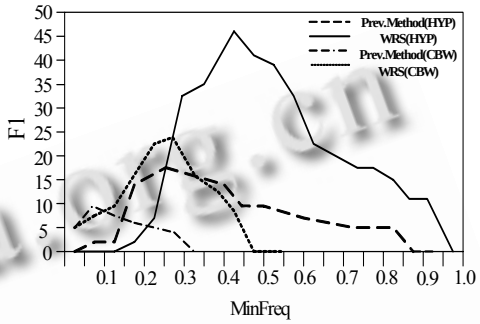


图 4　MinFreq 变化对综合参数 F1 的影响

## 3 结语

本文开发了一种基于 WEB 挖掘的个性化推荐系统 WRS，为了对用户访问模式进行更好的识别，使用了图形分割算法建模用户访问模式；并通过网页之间关联性建立了无向图以及对每条边赋了权重，然后进行聚类分析；为了有效快速的找到用户当前的会话，采用了最长公共子序列算法，经过搜索比较可以得出用户当前的访问模式。通过对此系统的实验评估，发现 WRS 有效的提高了推荐质量。

本文做了一些有效工作，但仍有不少问题需要进一步的研究，今后将进一步改进个性化推荐算法，并将提出的个性化推荐系统应用到电子商务网站系统中。

## 参考文献

1 戴军湘,李陶.WEB 日志挖掘技术研究及其在电子商务中的应用.科学技术与工程,2005,5(15):1081−1086.

2 张新香.WEB 日志挖掘在电子商务中的应用研究.计算机系统应用,2006,15(1):52−55.

3 Baraglia R, Silvestri F. Dynamic personalization of WEB sites without user intervention. Communications of the ACM, 2007,50(2):63−67.

4 Huan J, Kamber M. Data mining: Concept and techniques. San Mateo, CA: Morgan-Kaufmann, 2000.

5 何典,宋中山,刘济波.基于用户访问记录的 WEB 挖掘研究.

（下转第 119 页）

定义前件云模型为：

A1＝[-1，0.3，0.01]；A2＝[-0.5，0.3，0.01]；

A3＝[0，0.3，0.01]； A4＝[0.5，0.3，0.01]；

A5＝[1，0.3，0.01]。 (7)

定义后件云模型为：

B1＝[0.5，0.3，0.01]； B2＝[-1.5，0.3，0.01]；

B3＝[0，0.3，0.01]； B4＝[1.5，0.3，0.01]；

B5＝[-0.5，0.3，0.01]。 (8)

## 11 结语

本文全面论述了一种正态分布真随机数云模型发生器的实现方法，是对云模型研究的有益补充，为进一步使用真随机数云模型发生器做更深入的科学研究打下了坚实基础。

## 参考文献


1 郭弘,刘钰,党安红,韦韦.物理真随机数发生器.中国科学(科学通报),2009,54(23):3651–3657.

2 周丽娜,沈海斌,潘洋洋,董文箫.一种无记忆的真随机数发生器.电子器件,2008,31(3):945–947.

3 沈华韵,张鹏,王侃.改进线性同余法随机数发生器.清华大学学报(自然科学版),2009,49(2):191–193.

4 王欣,周童,王永生,喻明艳.一种基于混沌原理的真随机数发生器.微电子学与计算机,2009,26(2):135–139.

5 刘晓旭,曹林,董秀成.ATme ga128 单片机的真随机数发生器.单片机与嵌入式系统应用,2009,(11):71–73.

6 周童,周志波,喻明艳,叶以正.一种基于混沌的鲁棒低功耗真随机数发生器(英文).半导体学报,2008,29(1):69–73.

7 龙银东,敬岚,方正,乔卫民.用 VHDL 实现的 23 位快速浮点数加减法器.微计算机信息,2009,25(1-2):290–291.

8 王兆红,肖孟强,李燕,刘昕.类正态分布数据云模型的预测算法.计算机应用与软件,2009,26(9):78–79.

9 詹惠琴,古军,习友宝.正态分布随机 Petri 网的串并行结构化简.电子科技大学学报,2008,37(3):424–427.

10 徐燕娟,李众,张日勋.一维多规则正态云模型映射器的算法研究.科学技术与工程,2010,10(1):244–247.

11 杨金牛,李众,杨真荣.基于遗传算法的云模型控制器设计.计算机仿真,2009,26(3):175–178.

12 王飞燕,李峰,陈松贵.基于一维正态云模型的半脆弱文本水印.计算机工程与设计,2008,29(17):4578–4580.


（上接第 70 页）


计算机系统应用,2007,16(4):57–60.

6 Mobasher B, Cooley R, Srivastava J. Automatic personalization based on WEB usage mining. Communications of the ACM, 2000,43(8):142–151.

7 Coleho B, Martins C, Almeida A. Web Intelligence in Tourism: User Modeling and Recommender System. Web Intelligence and Intelligent Agent Technology(WI-IAT), 2010,236(1):619–622.

8 Herlocker JL, Konstan JA, Terveen LG, Riedl JT. Evaluating collaborative filtering recommender systems. ACM Trans. on Information Systems, 2004, 22(1):5–53.

9 Angulo C, Ruiz F J, Gonzalez L, et a1. Multi-classification by using tri-class SVM. Neural Processing Letters, 2006, 23(1): 89–101.

10 Adomavicius MG, Tuzhilin MA. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. on Knowl and Data Eng, 2005,17(6):734–749.