

# 基于兴趣度的 Web 页面关联规则的研究

陈永平, 刘 俞, 苏 新

(安徽工业大学(南校区), 安徽 马鞍山 243002)

**摘 要:**随着互联网上的信息迅速增长,如何快速准确地寻找到信息越来越受到人们的重视。文中给出了几种计算用户兴趣度的方法,并利用其中一种计算用户兴趣度的方法,论述了基于兴趣度的 Web 页面关联规则。论述了关联规则和一般的 Apriori 算法,并利用了“壹支持数下 K-关联规则”,对一般的 Apriori 进行了改进,主要是将兴趣度用于 Apriori 算法中。实验结果证明,该方法用于在网上寻找用户感兴趣的信息具有较好的准确率。

**关键词:**兴趣度;Web 挖掘;关联规则;Apriori 算法

**中图分类号:**TP301.6

**文献标识码:**A

**文章编号:**1673-629X(2008)05-0086-04

## Research of Web Page Association Rules Based on Interesting

CHEN Yong-ping, LIU Yu, SU Xin

(South Branch of Anhui University of Technology, Ma'an Shan 243002, China)

**Abstract:** As the rapidly growing of information on Internet, how to rapidly and precisely find the information have been more and more noticed by the people. Give some methods for user interesting computing, and use one of the methods for user interesting computing. Prove the association rules based on the interesting Web. Also prove the association rules and the usually Apriori algorithm, and use "K association rules on one-support count", set idea and some betterment algorithms for Apriori which based on frequency item of mining association rules, mainly used the interesting to the algorithms for Apriori. Experiment shows that this method can provide a better precision for people find the interesting information.

**Key words:** interesting; Web mining; association rules; Apriori algorithm

## 0 引 言

随着互联网的飞速发展,网上的资源空前丰富。但是数据资源中蕴含的知识至今未得到充分的挖掘和应用。数据挖掘技术的出现和发展为解决这个问题带来了希望。数据挖掘就是从大量的数据中提取或“挖掘”知识。将数据挖掘和 Web 结合起来就产生了 Web 挖掘。Web 挖掘就是利用数据挖掘技术从 Web 文档和 Web 活动中发现和抽取人们感兴趣的、潜在的有用模式和隐藏的信息。它实现对 Web 存取模式、Web 结构和规则,以及动态的 Web 内容查找。Web 用户访问信息挖掘(也称为 Web 日志挖掘)是从存取模式中获取有价值的信息和模式的过程,就是对用户访问时留下的访问记录进行挖掘<sup>[1]</sup>。每个用户都有各自的访问目的,因而具有不同的访问序列。如果当前用户已有一个访问序列,那么具有类似访问序列的其他用户的

下一次访问可以为该用户提供推荐。

根据这一点,文中提出了一种基于兴趣度的关联规则的研究,目的是以用户访问过的序列来预测用户将要访问的信息。

## 1 数据预处理

数据预处理是数据挖掘处理过程的第一步。一般地,数据挖掘预处理过程包括数据清洗、数据集成和转换、数据归纳以及数据选择等步骤<sup>[2]</sup>。通过这些处理,剔除了原来数据集中的脏数据和噪音数据,补充其中丢失的数据,将多个异地、异构的数据转换成数据挖掘所需要的数据源的形式,使数据挖掘在比较规范的数据源中进行,从而减少挖掘算法的数据处理量,在提高挖掘效率的同时,也提高了知识发现的起点和发现知识的有效性。

对不同的应用领域,数据挖掘的预处理、发现规则和模式分析等过程的处理也不尽相同。而 Web 挖掘中数据预处理主要包括数据净化、用户识别、会话识别及路径补充<sup>[3]</sup>等几个步骤。

收稿日期:2007-08-26

基金项目:安徽省自然科学基金资助项目(KJ2007B245)

作者简介:陈永平(1969-),男,讲师,硕士,研究方向为数据挖掘、算法设计。

## 2 兴趣度

兴趣度就是用户对某一网络结点或某一Web网页的兴趣度强弱。它反映了用户对网络结点或Web网页的喜爱程度。可以通过用户对某一网页的浏览行为来计算和表示用户的兴趣度。目前,用来计算和表示用户的兴趣度主要有以下几个公式:

公式1:定义USER对事务页PAGE的兴趣度F(USER, PAGE)如下<sup>[4]</sup>:

$$F(\text{USER}, \text{PAGE}) = \frac{\text{user 浏览 page 所用的时间}}{\text{user 的总浏览时间}} \times \frac{\text{page 字节数}}{\text{路径中总字节数}} \quad (1)$$

公式2:定义USER对事务页PAGE的兴趣度F(USER, PAGE)如下<sup>[5]</sup>:

$$F(\text{USER}, \text{PAGE}) = \frac{\text{user 浏览 page 所用的时间}}{\text{page 中字符数目}} \times \text{路径因子} \quad (2)$$

其中路径因子可以根据PAGE在用户会话有意义的访问路径P中的深度设定。例如:若PAGE是P的终点,则PAGE的路径因子设为大于1的值,否则,设为1。

公式3:F(USER, PAGE)为用户USER对Web事务页PAGE的兴趣度,用多元线性回归模型来描述用户兴趣与用户浏览Web事务页时间和翻页/拉动滚动条次数(浏览时间和拉动滚动条次数为用户浏览兴趣事务页时的两种主要行为),其公式如下<sup>[6]</sup>:

$$F(\text{USER}, \text{PAGE}) = AX_1 + BX_2 + C \quad (3)$$

其中: $X_1$ 表示浏览时间, $X_2$ 表示翻页/拉动滚动条次数, $A, B, C$ 为一组常数(站点类型不同有不同的值,为一经验值)。

以上三公式各有优缺点,文中采用公式3来计算用户的兴趣度。

## 3 关联规则

### 3.1 关联规则的定义

关联规则的形式化描述为:设  $I = \{i_1, i_2, \dots, i_m\}$  是  $m$  个不同项的集合, $D$  是针对  $I$  的事务的集合,其中每一事务  $T$  包含若干项  $i_1, i_2, \dots, i_k \subset I, T \subseteq I$ , 有一个标识符 TID。关联规则表示为  $X \Rightarrow Y$ , 其中  $X, Y \subset I$ , 并且  $X \cap Y = \emptyset$ ,  $X$  称为规则的前提或前项, $Y$  是结果或后项。

项集(itemset):一些项的集合,在项集中的项数称为项集的长度,包含  $k$  个项的项集称为  $k$ -项集;

支持度(support):对于  $X \subset I$ , 若  $D$  中包含  $X$  的事务个数为  $s$ ,  $D$  中事务总数为  $n$ , 则  $\text{support}(X) = s/n$ ;

置信度(confidence):定义为  $\text{confidence}(X \Rightarrow Y) = \text{support}(X \cup Y) / \text{support}(X)$ ; 对于关联规则

$X \Rightarrow Y, \text{support}(X \Rightarrow Y) = P(X \cup Y), \text{confidence}(X \Rightarrow Y) = P(Y/X)$ ;

阈值:最小支持度 minsup, 最小置信度 minconf。

频繁项集:指项集的出现频率  $\geq \text{minsup}$ , 频繁  $K$ -项集的集合通常记作  $L_K$ 。所有的大项集可表示为:  $L = \bigcup_{k=1}^{\max K} L_K$ , 其中  $\max K$  是最大项集的长度。

非频繁项集:是指支持率  $< \text{minsup}$  的项集。

性质1:频繁项集的子集也是频繁项集。

性质2:非频繁项集的超集也是非频繁项集。

性质3:如果  $X, Y$  是项集, 并且  $X$  是  $Y$  的子集, 则  $\text{support}(X) \geq \text{support}(Y)$ 。

关联规则的问题就是找出这样一些规则: 它们的 support 和 confidence 分别大于等于指定的 minsup 和 minconf。

### 3.2 关联规则挖掘算法

关联规则的挖掘算法主要可以分为以下几个方面<sup>[7]</sup>:

1) 利用频繁项集向下封闭的性质(Apriori 性质)的 Apriori 系列算法, 主要有: 经典关联规则挖掘算法 Apriori 算法; 对事务项集进行重组预选的 AprioriTid 算法; 用 Hash 表进行事务项集重组的 DHP 算法等。

2) 利用对事务数据库划分来提高效率算法。

3) 利用各种数据结构进行的数据挖掘的算法。

4) 遗传算法。

## 4 基于兴趣度的页面关联规则

### 4.1 相关理论

前面已经介绍了关联规则的相关知识, 目前它已经应用于 Web 个性化、网上问答系统、网上图书馆等多个领域, 并取得一定的效益。应用 Web 使用挖掘技术, 从 Web Log 文件中挖掘关联规则, 可以得到如下一些规则。

访问了 a1 页面的用户中, 有 40% 的用户又访问了 c1 页面。

访问了 c2 页面的用户中, 有 70% 是从 a2、a3、a4、a5 进入系统的。

从 a6 页面进入系统的用户中, 有 60% 又访问了 c3 页面。

以上 Web Log 文件中发现的规则中,  $a_i (i = 1, 2, 3, 4, 5, 6)$  和  $c_j (j = 1, 2, 3)$  分别为关联的前项和后项, 40%、70% 和 60% 为规则的置信度。根据关联规则的表示形式, 可以表示为:

$a1 \Rightarrow c1, \text{confidence} = 40\%$

$a2a3a4a5 \Rightarrow c2, \text{confidence} = 70\%$

$a6 \Rightarrow c3, \text{confidence} = 60\%$

在关联规则的应用中,所应用的关联规则具有如下的特点<sup>[8]</sup>:

规则后项代表用户访问过的一个页面,其长度为 1,即  $1-size$ ;

规则前项代表用户在访问后面之前所浏览的页面序列,其长度是该用户浏览过的页面数目,可称其为会话窗口数目。

经过分析表明,为了保证推荐具有较高的覆盖率和准确率,其应用的关联规则应具有  $1-size$  的后项和适当的前项。同时,不同领域、不同数据集以及不同的推荐对于支持度设置的要求也将不一定相同。为解决支持度阈值的设置,提出了“壹支持数”的概念<sup>[9]</sup>,并引出了一些相关概念和定理。

**定义 1** 壹支持数即寻找频繁项集时设定的支持数阈值为 1。

**定理 1** 采用壹支持数可以获得数据集中全部项集。

根据项集定义,满足最小支持数阈值的项组合为频繁项集,而根据上面定义,一个项集只要在任意记录中出现一次,即可成为频繁项集,这样,频繁项集包含了所处理数据集中的全部项集。

**定义 2** 壹支持数下  $k$  关联规则为:根据壹支持数阈值寻找频繁项集而在生成关联规则时,对于所有前项相同的规则,只选择其中  $K$  个支持数最大者作为推荐规则。

**定理 2** 在壹支持数下  $k$  关联规则条件下,对于  $m$  个  $i-size$  ( $2 \leq i \leq n+1$ ) 频繁项集  $item_{i1}, item_{i2}, \dots, item_{im}$ ,若它们均包含某个  $(i-1)-size$  频繁项集中的各个项,则仅选择其中支持数最大的  $k$  个项集生成前项长为  $(i-1)-size$  的推荐关联规则即可。

#### 4.2 基于兴趣度的页面关联规则

目前大部分关联规则挖掘算法仍采用支持度和置信度阈值作为寻找频繁项集和生成规则的依据,在应用关联推荐算法中,虽然在推荐时也对那些待推荐的页面加权,加权时考虑页面的访问时间等因素,但这是在关联规则生成之后,推荐时如有多个待推荐页面,而某页面的访问频繁度高,并不能绝对地说明用户对该页面感兴趣。

如页面 A 具有 50% 的支持度,0.3 的用户平均访问兴趣度,页面 B 具有 10% 的支持度和 0.7 的用户平均访问兴趣度,如果给定的支持度阈值为 20%,B 页面将不会成为频繁页面,更不可能生成关于 B 的关联规则,但是 0.7 的兴趣度和 0.3 的兴趣度对比却说明用户对 B 页面的兴趣远远大于 A,由于 B 页面的支持度小不能成为频繁页面,也不能生成关于 B 的关联规则,

更不可能被加权,因而不能被推荐,这在很多情况下是不合理的。针对这种情况,文中采用了将兴趣度用于关联规则挖掘中,从而能更为合理得到频繁项集,以实现较准确的推荐。

前面所述的预处理后的会话文件,是对相关 Log 文件进行了数据净化、用户识别、会话识别及路径补充后的文件,该文件是一个具有  $n$  个页面的集合  $P$  和具有  $m$  个会话的集合  $S$ ,即  $P = \{p_1, p_2, \dots, p_n\}$  及  $S = \{s_1, s_2, \dots, s_m\}$ ,这里,每个  $s_i \in S$  为  $P$  的子集,而每个  $s$  可以表示为一个  $h$  长的序列,序列中元素为一个有序对  $s = \langle (p_{i1}, ip_{i1}), (p_{i2}, ip_{i2}), \dots, (p_{im}, ip_{im}) \rangle$ ,这里  $p_{ij} \in P$  ( $j = 1, 2, \dots, h$ ),  $ip_{ij}$  为  $p_{ij}$  的兴趣度的值。本节的预处理是对上述会话文件的进一步处理,它包括页面访问的兴趣度的计算和支持的生成两部分。在一般的关联规则挖掘中,对于每个  $s \in S$ ,如果一个页面  $p \in P$  及  $p \in S$ ,则  $p$  的支持数增 1,即一个频繁页面序列的支持数表示包含该序列的会话记录数,也即每个包含该序列的会话记录对该序列支持数的贡献为 1,而在本节中,将一个页面的原支持数“1”与对该页面的访问兴趣度  $ip$  结合后,采用如下公式生成新的支持度  $\varphi(p)$ 。

$$\varphi(p) = \alpha + \beta \cdot ip \quad (4)$$

对于每个  $s \in S$ ,如果一个页面  $p \in P$  及  $p \in S$ ,则  $p$  的支持数增加  $\varphi(p)$ 。这里  $\alpha$  和  $\beta$  分别为页面原支持数和页面访问的兴趣度在新支持度  $\varphi(p)$  中所占的比例,即为:  $\alpha \times 1 + \beta \times ip$ ,要求  $0 \leq \alpha, \beta \leq 1$ ,且  $\alpha + \beta = 1$ , $\alpha, \beta$  值一般设置为 0.5 时能够得到最大的综合测度。

**关联规则挖掘算法:**

该算法主要是对 Apriori 算法进行改进,在 Apriori 算法,如果一个频繁页面序列存在于一个会话记录中,则执行  $c.conut = c.conut + 1$ ,而在改进后的算法中则执行  $c.count = c.count + \varphi(p)$ 。改进后的 Apriori 算法描述如下:

输入:事务数据库 D;最小支持度阈值 minsup;

输出:D 中的频繁项集 L

**算法描述:**

$L_1 = \text{find\_frequent\_1\_itemsets}(D)$ ; //生成 1—size 频繁项集的集合  $L_1$

For( $k=2; L_{k-1} \neq \emptyset; k++$ )

{

$C_k = \text{apriori\_gen}(L_{k-1})$ ;

//由  $K-1$  频繁项集生成  $K$  候选集

for each transaction  $tD$

{//搜索 D 以计算支持数

```

 $C_i = \text{subset}(C_k, t)$ 
//从事务  $t$  获得包含在  $C_k$  候选集中的子集,该子集的项集都出
现在事务  $t$  中
for each candidate  $c \in C_i$ 
if  $c \subseteq D$  then
 $c.\text{count} = c.\text{count} + \varphi(p)$ ; //计算候选集中包含在  $C_k$  中每一
项的支持数
}
 $L_k = \{c \in C_k | c.\text{count} \geq \text{minsup}\}$ 
}
return  $L = \bigcup L_k$ ;

```

对于一个  $i$ -size 频繁页面序列 ( $i \geq 2$ ),  $\varphi$  应用公式(5)进行计算:

$$\varphi(pj1, pj2, \dots, pji) = \frac{\sum_{K=1}^i \varphi(pjk)}{|pj1, pj2, \dots, pji|} \quad (5)$$

生成的关联规则的算法与上面改进的 Apriori 算法相同。

在关联规则应用中应具有这样的形式:  $p_{a1}p_{a2} \dots p_{ai} \Rightarrow p_c$  ( $i = 1, 2, \dots, n$ )。该规则可以解释为用户的访问习惯通常是访问了  $p_{a1}, p_{a2}, \dots, p_{ai}$  后访问  $p_c$ 。根据这一规则,当一个 Web 站点的用户访问了  $p_{a1}, p_{a2}, \dots, p_{ai}$  页面后,该站点的推荐引擎就会为该用户推荐  $p_c$  页面。

## 5 实验

在这里采用安徽工业大学网站的处理数据,对基于兴趣度的页面关联规则的推荐算法在覆盖率和准确率进行测试和比较<sup>[9]</sup>。该数据集具有 498 个 URL, 8789 个会话记录。选择其中前 300 个访问记录。经预处理后,将数据集的 3/5 作为训练集,进行 Web 挖掘以生成推荐的关联规则,其余 2/5 作为测试集,进行测试。如图 1 和图 2 所示。图中给出了在考虑兴趣度情况下的覆盖率和准确率随最小支持度阈值的变化情况。

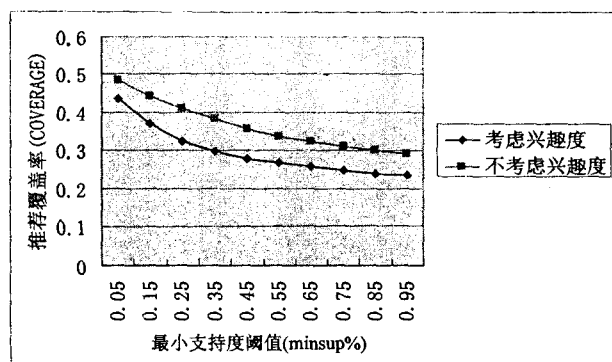


图1 推荐覆盖率随最小支持度阈值的变化情况

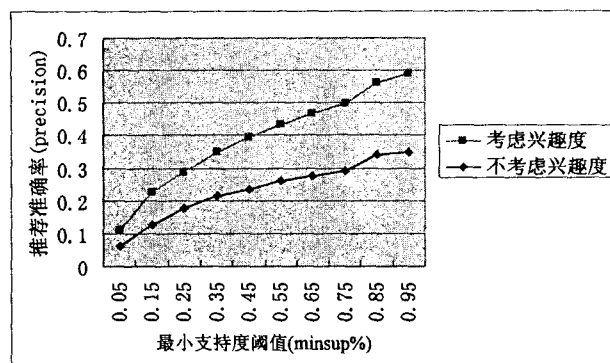


图2 推荐准确率随最小支持度阈值的变化情况

由上面两图可知,基于兴趣度的关联规则在推荐时覆盖率较低,这是由于产生的规则数目较少,因而可推荐页面较少,但是,基于兴趣度的页面关联规则进行推荐时具更高的准确率(在这里  $\alpha$  和  $\beta$  的取值都是 0.5)。

## 6 结束语

由于关联规则在 Web 个性化、个性化搜索引擎、网上图书馆、自动问答系统、情报检索等方面的应用,使得关联规则的研究受到越来越多人的重视。文中在介绍关联规则相关知识的同时,介绍了基于兴趣度的 Web 页面关联规则的研究,并给出基于兴趣度的关联规则挖掘算法。通过实验证明基于兴趣度的关联规则挖掘在推荐准确率上要比不考虑兴趣度的高。这也说明了将该方法应用于实践中将会得到较好的效果。

## 参考文献:

- [1] Rakotonirainy A, Bond A, Indulska J, et al. A simple component architecture framework [C]//Technology of Object - Oriented Languages, 2000, TOOLS 33, Proceedings, 33rd International Conference on. St. Malo, France: [s. n.], 2000: 359 - 370.
- [2] Han J, Kamber M. Data Mining: Concepts and Techniques [M]. USA: Morgan Kaufmann Publishers, 2001.
- [3] Cooley R, Mobasher B, Srivastava J. Data Preparation for Mining World Wide Web Browsing Paterns [J]. Knowledge and Information System, 1999, 6(1): 5 - 32.
- [4] 张新香. Web 日志挖掘在电子商务中的应用研究 [J]. 计算机系统应用, 2006(1): 52 - 55.
- [5] 张 莹. 从商务网站用户行为数据提取用户兴趣 [J]. 潍坊学报, 2005, 5(4): 21 - 23.
- [6] 周晓兰, 王随平. Web 文本挖掘中用户兴趣模型的建立和更新 [J]. 湘潭师范学院学报: 自然科学版, 2006, 28(3): 33 - 36.
- [7] 周 涛, 陆惠玲. 关联规则挖掘算法研究 [J]. 齐齐哈尔大

(下转第 93 页)

调用了处理此种省略情况的规则:“动词搭配分析-失败分类分析-2-取前一句主语1a”。

此规则的原理为:取前一句的第一个成功事件的主语作为新分析事件的主语。即处理的是主语承前省略的情况。

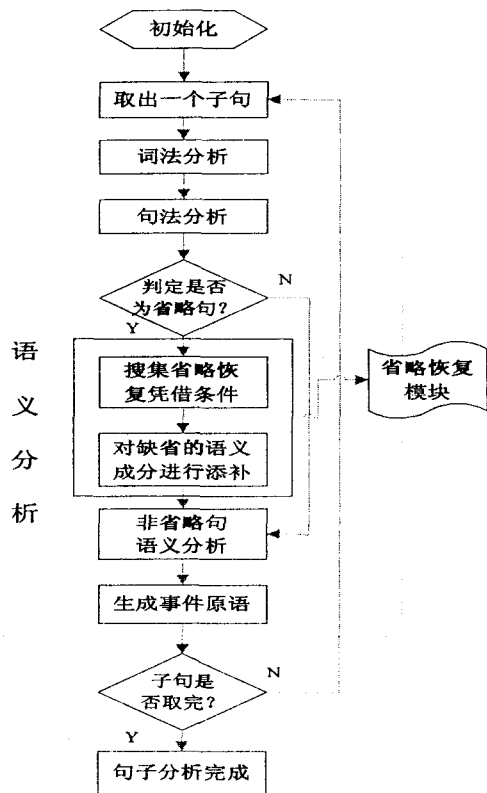


图2 语句分析处理流程

运行完成后的语义理解结果(生成包含如下文字的文档):

“(动态事件(标识 标识%0112110100)(事件名 线段相交事件)(时间 nil)(地点 nil)(结果 交点: G 线段相交关系:  $AB \cap EF$ )(施动对象 AB)(受动对象 EF)(已理解)(所在句 标识 %0112100000))”

从分析结果可以看出,线段 EF 的主语已经根据找前一字句(线段 AB 垂直线段 CD)的主语(AB)添补完成,施动对象为线段 AB,受动对象为线段 EF。

## 6 结束语

实验室省略恢复模型仅实现了部分简单的省略情形的恢复研究,对于需要通过语境来逻辑判断缺省部分的情形还没有完善。例如语句:王老汉五十岁那年

才添了儿子,( )现在都上小学了。通过系统语用分析发现缺少主语,如果按照系统取前一字句成功主语的规则,后一子句添补的主语就是老汉。可能语义搭配分析上能过去,但在逻辑上不符合。这就需要在篇章分析时加入判断,从而语境上判定主语(缺省部分)的内容——即和语用学<sup>[5]</sup>相结合。

随着系统模型的逐渐完善,相信对于占省略情形最大比例的主语省略,定语、宾语省略以及顶针式省略都能高准确率地实现恢复。

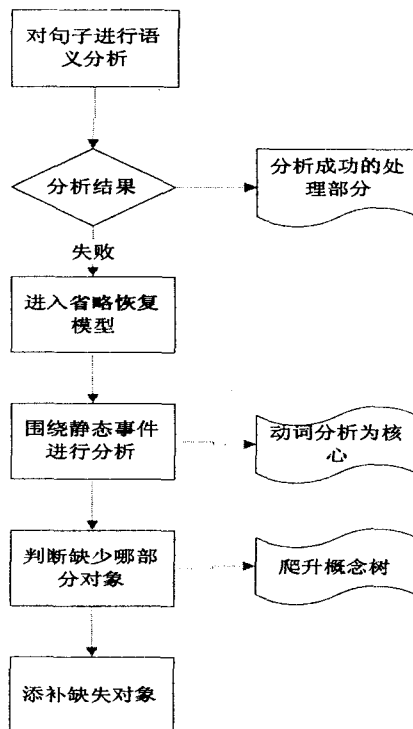


图3 省略恢复模型

## 参考文献:

- [1] 钱世凤. 省略界定综述[J]. 语文学刊: 高教版, 2007(1): 119-122.
- [2] 殷 鸿. 基于概念模型的省略恢复研究[J]. 计算机工程, 2007(24): 196-199.
- [3] 赵世举. 关于汉语省略句的判定标准问题[J]. 中南民族学院学报: 哲学社会科学版, 1999(4): 98-100.
- [4] 党 建. 数学领域集体词结构形式化处理研究[J]. 计算机技术与发展, 2007, 17(5): 121-124.
- [5] 赵礼彬. 基于领域的语用研究和实现[J]. 计算机技术与发展, 2007, 17(6): 49-52.

(上接第89页)

学学报, 2004, 20(3): 58-62.

- [8] Lin W, Alvare S A, Ruiz C. Efficient adaptive support association rule mining for recommender systems[J]. Data Mining

and Knowledge Discovery, 2002, 6(1): 83-105.

- [9] 闫 莺, 王大玲, 于 戈. 支持个性化推荐的 Web 页面关联规则挖掘算法[J]. 计算机工程, 2005, 31(1): 79-81.