基于概念层次树的用户特征挖掘技术*

田凤珍,韩宪忠,陈 晨,王克俭
(河北农业大学 信息科学与技术学院,河北 保定 071001)

摘 要:将面向属性的归纳方法应用到网上书店中,通过概念层次技术从用户的注册信息中归纳出用户的访问需求,从而实时主动地为用户提供个性化服务。实验证明该方法对研究用户的兴趣爱好有意义。

关键词:用户特征;属性归纳;概念层次树;网上书店

中图分类号:TP391 **文献标识码**:A **文章编号**:1674-7720(2011)05-0095-03

Mining user characteristics based on the concept hierarchy tree

Tian Fengzhen, Han Xianzhong, Chen Chen, Wang Kejian
(College of Information Science and Technology, Agricultural University of Hebei, Baoding 071001, China)

Abstract: The paper applied the approach of attribute-oriented induction to online bookstore, summarize access requirements from the user registration information through the concept hierarchy technology, to actively provide real-time personalized service. The experiment results show that the method is meaningful to study the user interests.

Key words: user characteristics; attributes induction; concept hierarchy tree; online bookstore

电子商务的迅猛发展使得越来越多的用户把注意力转移到网络上,但是由于Web信息量及复杂度的迅速上升,直接导致用户面对庞大的网络时无从着手。因此,研究如何使用户更加方便快捷地访问网站,如何在短时间内更有效地获取用户需求的信息,成为当前各个网站构建与优化需要重点考虑的问题。

网站为用户提供个性化服务是有效解决以上问题的方法之一。个性化服务就是网站通过收集和分析用户的注册信息及访问行为等知识,预测用户未来的网页请求、了解用户的兴趣爱好、分析用户的访问模式,根据用户的个性化需求,为用户实时、主动提供所需求的信息页面。通过网络提供的个性化服务可以有效地解决用户"信息过载"和"信息迷失"的困扰。一方面可以方便用户使用,提高用户的访问效率,满足用户的个性化需求;另一方面对企业在改善顾客关系、培养顾客忠诚以及增加网上销售方面也具有明显的效果。

目前在个性化服务中的研究主要集中在从用户的访问行为中挖掘特征规则。Mobasher[1]提出了一种基于Web使用挖掘的个性化服务体系结构,通过使用聚类和关联规则发现方法为用户提供个性化服务;张成[2]等提出了一种基于OWL-S的服务挖掘算法,通过计算服务关键字权重得出服务的相识度,来分析服务之间的匹配,从而定位所需服务,在一定程度上提高了服务的性能,但挖掘用户注册信息中特征规则的研究居少;卢明等[3]提出一种使用属性表的快速概念聚类算法,通过构造一张属性表对前缀树进行剪枝,概念聚类的过程仅在一些有效的子空间中执行。

本文主要研究如何从用户注册信息中挖掘出用户的特征规则。结合网上书店,应用面向属性的归纳AOI(Attribute Oriented Induction)方法,从关系数据库的用户注册信息中挖掘与用户购书行为有关的特征规则,从而推断同类用户将来的购书需求,并为调整网站结构及个性化服务提供依据。

1 网络用户特征的挖掘方法

对于网站来说,将网络用户分为注册用户和非注册用户。挖掘网络用户特征主要从两方面研究,分别为用户注册信息特征(从用户的注册信息中归纳出的特征)和用户行为特征(从用户在网站的浏览行为中归纳的特

*基金项目:河北省自然科学基金资助项目(F2009000653),河北省科学技术厅计划资助项目(072135126),河北省教育厅计划资助项目(Z2009122)

征)。对于非注册的用户,其基本信息获取比较困难,故欲不考虑这部分用户。

面向属性的归纳方法主要是根据用户的属性数据概化出用户的特征,从而得知用户的需求,被广泛地应用于特征规则、多层规则和分类规则的挖掘。特征规则(Characterization Rule)描述的是目标数据集中大部分数据所共有的特征。挖掘方法有概念描述(Concept Description)方法和数据泛化(Data Generalization)方法。概念描述是对某类对象的内涵进行描述,并概括这类对象的有关特征;数据泛化是一个将数据集中的属性从较低的概念层抽象到较高的概念层的过程。实现数据泛化的方法有数据立方体(OLAP)方法和面向属性方法两种:(1)OLAP方法是通过一系列分析处理过程将数据集中的数据以不同的数据组织方式和可视化的形式呈现给用户;(2)面向属性归纳方法则采用概念分层的思想,通过以高层概念替换低层数据来实现泛化[4]。

概念层次结构是表示抽象知识的重要手段,把原始数据泛化到较高层次,实现在不同概念层次上对数据的抽象。面向属性归纳方法中用来进行概念泛化的技术称为概念层次技术,用概念层次树来表示用于泛化的背景知识,实现具体与抽象概念之间的转化。概念层次树是将数据库中记录的属性字段根据一定的抽象程度进行归类合并而形成的层次结构。面向属性归纳方法利用概念层次技术进行概念提升,得到高度概括的表,再进而将它转换成用户的特征需求,为用户个性化服务提供依据。

目前挖掘网络用户行为特征主要应用Web数据挖掘技术,通过挖掘用户访问Web时在服务器上留下的日志记录,来发现用户访问Web页面的模式。主要有用户聚类和网页聚类两种聚类技术用来挖掘用户行为特征。用户聚类主要是把所有用户划分成许多组,具有相似浏览模式的用户分在一组。网页聚类则可以找出具有相关内容的网页组,根据用户的询问或过去所需信息的历史来生成静态或动态网页,从而向用户推荐相关的超链接。

综上所述,从用户注册信息和行为两方面挖掘出的特征规则都能反映出用户的兴趣爱好、个人需求信息等,可以同时为网站的构建及优化提供依据,从而达到为用户提供个性化服务的目的。

2 挖掘网上书店的用户注册信息中的特征规则

实验中服务器系统为Windows Server 2003,版本为Enterprise Edition。构建了一个网上购书网站,用户的注册信息以记录的形式存储在关系数据库中。选取数据库中的记录,通过概念层次技术,挖掘出与用户有关的特征规则。

用户的注册信息包括用户名、性别、年龄、职业、教育程度、收入、喜欢的书等基本信息。其中职业的分类是按照国家标准分为八大类,分别为国家机关及企事业单位负责人、专业技术人员、办事人员及有关人员、商业和服务业人员、农林等业的生产人员、生产及运输设备操作人员、军人、其他八大类。教育程度分为小学、初中、职中、高中、中专、大专、本科、研究生及以上八类。收入由四个分界点分为五个不同的层次。喜欢的书分为小说类、历史、人文社科、计算机类、管理学、其他六类。实验中要求用户在注册网站时需要选择基本信息的相关选项,以下是挖掘用户注册信息中的特征规则的步骤。

2.1 建立概念层次树

根据关系数据库中的数据,为用户的每个属性构建概念层次树,使具体的属性值概化为抽象的知识并归类合并,实现在不同概念层次上对数据的抽象。概念层次树是通过树结构的形式,将具体的属性值分组,然后按照背景知识逐级提升概念。每个独立节点表示一个基本概念,它可能是一个属性的简单组,也可能是若干属性形成的复合组。概念层次树的节点可以是同一属性的不同抽象度的汇聚点,也可以是由一个概念包含的多个子概念。网站中用户的性别、年龄、职业、教育程度、收入、喜欢的书等基本信息都不同程度上对用户的兴趣爱好产生一定的影响,所以需要把这些基本信息的概念层次树构建出来,其中"喜欢的书"这一基本信息是两层的概念层次树,与"职业"的相同。

图1~图4所示分别是性别、年龄、职业、教育程度、收入的概念层次树。

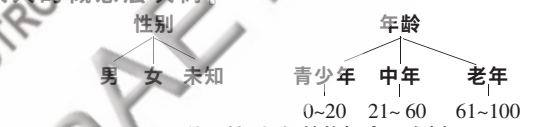

图1 关于性别、年龄的概念层次树

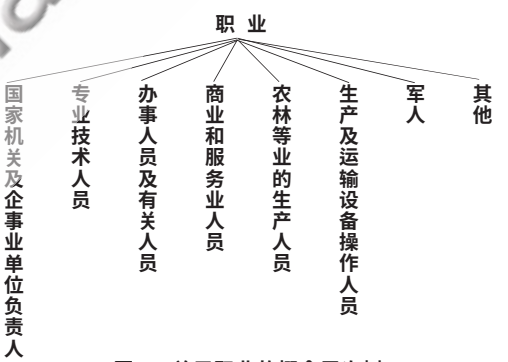

图2 关于职业的概念层次树

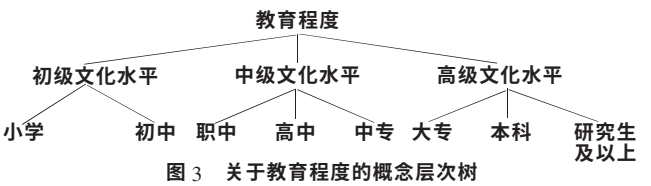

图3 关于教育程度的概念层次树

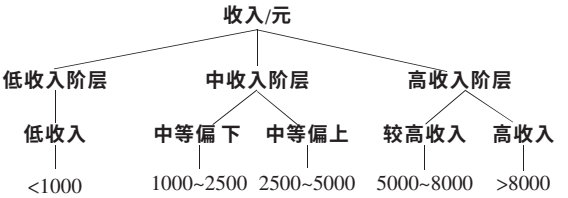

图4 关于收入的概念层次树

2.2 描述概念层次的数据库表

为了将概念层次树存入数据库，在表 1 中列出了描述概念层次的数据表。将概念层次树中的属性信息映射到数据库表中，表中给出了层次编码、概念节点的名称、层号、属性标志和与概念相对应的属性取值的区间。概念层次树中叶节点为第 0 层，叶节点的父概念所在节点为第 1 层，以此向上类推。如果概念层次树是两层，在数据表中层次编码用两位数字；如果是三层的，编码用三位数字。性别的属性标志是 1，年龄的属性标志是 2，教育程度的属性标志是 3，收入的属性标志是 4，以此类推。

表 1 描述概念层次的数据表

层次编码	名 称	层号	属性标志	属性取值
10	性别	0	1	
11	男	1	1	x=1
12	女	1	1	x=2
13	未知	1	1	x=3
⋮	⋮	⋮	⋮	⋮
300	教育程度	0	3	
310	初级文化水平	1	3	
320	中级文化水平	1	3	
330	高级文化水平	1	3	
311	小学	2	3	x=1
312	初中	2	3	x=2
⋮	⋮	⋮	⋮	⋮
400	收入/元	0	4	
410	低收入阶层	1	4	
411	低收入	2	4	1 000 以下
420	中收入阶层	1	4	
421	中等偏下	2	4	1 000~2 500
422	中等偏上	2	4	2 500~5 000
430	高收入阶层	1	4	
431	较高收入	2	4	5 000~8 000
432	高收入	2	4	8 000 以上

2.3 特征规则挖掘的处理过程

(1)由概念层次描述的数据建立数据库表进行组合条件计算，具体包括基于单一属性的概念提升和生成基于多属性的关联条件；

(2)进行类组(基本概念或复合概念相对应的数据子集)数据计算，内容包括生成数据库子集、对类组记录进行排序及数据统计等。排序时可以计算高收入阶层占总购书的比例或者高文化水平占总购书的比例，并以此作为排序的依据。

用户注册成功后，根据用户的信息映射到数据库表中，按照得出的规则特征为用户提供感兴趣的图书及网页等。

3 实验结果与分析

通过购书网站的关系数据库中选取 2010.04~2010.05 期间的 3 625 条记录，选取“教育程度”、“收入”、“喜欢的书”三个属性进行分析。首先按照 2.1 节建立这三个属性的概念层次树，使具体数据值抽象化。然后按 2.2 节将属性信息映射到数据库表中，经过概念提升和类组计算后，得出以下一些特征规则：

教育程度=初级文化水平 & 收入=低收入阶层 & 喜欢的书=小说类+其他 & 购书→1.15%

教育程度=初级文化水平 & 收入=中收入阶层 & 喜欢的书=人文社科+管理学 & 购书→5.23%

教育程度=中级文化水平 & 收入=中收入阶层 & 喜欢的书=人文社科+管理学 & 购书→2.36%

教育程度=中级文化水平 & 收入=高收入阶层 & 喜欢的书=计算机类+管理学 & 购书→7.56%

教育程度=高级文化水平 & 收入=中收入阶层 & 喜欢的书=人文社科 & 购书→2.17%

教育程度=高级文化水平 & 收入=中收入阶层 & 喜欢的书=计算机类+管理学 & 购书→6.28%

教育程度=高级文化水平 & 收入=高收入阶层 & 喜欢的书=人文社科 & 购书→3.34%

教育程度=高级文化水平 & 收入=高收入阶层 & 喜欢的书=计算机类+管理学 & 购书→8.13%

在只考虑这三个属性的情况下，得出以上特征规则。但是这些购书比例会受到其他因素的影响而有所不同，例如用户并不是按实际情况选取属性信息、或者用户临时需求要选取一类书等。由于用户的注册信息在一定程度上对购书结果产生了影响，所以研究用户注册信息的特征对了解用户兴趣爱好并提供个性化服务是有意义的。

面向属性归纳的方法能够根据用户的注册信息，通过概化技术，初步归纳出用户的兴趣爱好。该方法应用到网上书店中，为用户的个性化服务提供了依据。同时为以后更准确地提供给用户需求的信息也提供了研究依据。

参考文献

[1] MOBASHER B.A Web personalization engine basedon user transaction clustering [C] //In Proceedings of the 9th Workshop on Information Technologies and Systems(WITS'99), December 1999.

[2] 张成，张璟.一种服务挖掘算法的研究与实现[J].计算机工程与应用，2010，46(4)：117-119.

[3] 卢明，胡成全，齐红，等.一种使用属性表的快速概念聚类算法[J].复旦学报，2004，43(5)：823-826.

[4] 孙华梅，郭茂祖，焦杰，等.一种新的面向属性归纳中概念层次技术研究[J].管理科学学报，2004，7(1)：65-72.

(收稿日期：2010-09-13)

作者简介：

田凤珍，女，1983 年生，在读研究生，主要研究方向：计算机网络与数据库。

韩宪忠，男，1964 年生，教授，主要研究方向：网络信息处理，信息集成。