基于关联规则数据挖掘的研究及应用*

张延萍
(盐城工学院 院长办公室,江苏 盐城　224003)

摘　要:在研究了经典的关联规则算法 Apriori 之后,提出了类 Spriori 的数据挖掘算法分析学生频繁访问的页面路径,用以提供有用的信息给网络课程设计者以及授课老师,来解决学生在网络学习过程中产生的"信息迷航"问题。

关键词:关联规则;Apriori;数据挖掘;远程教学
中图分类号:TP311.1　　文献标识码:A　　文章编号:1671-5322(2007)01-0047-03

随着计算机的普及和 Internet 网的推广,计算机网络技术为个性化学习提供了良好的技术支持。经过多年的教学实践应用,在现有的远程教育站点上积累了大量有用的信息,然而这些信息存储分散,记录凌乱,数据庞大,如:学生注册信息,登录信息,浏览路径信息,答疑信息,作业信息,测试信息,交流信息,学习状态信息,学习进度信息等大量资源,如何利用这些资源建立一个智能化、个性化的远程教育环境,是现代远程教育技术发展中的一个关键问题。

近来随着数据挖掘技术的发展与成熟,人们逐渐利用数据挖掘技术从大量的已有累计数据中发现有利用价值的信息。现在数据挖掘技术已广泛应用在生物医学和 DNA 数据分析、金融数据分析、电子商务、电信业中,并取得很大的成效,但在远程教育领域中,数据挖掘技术并没有得到应有的重视。随着远程教育的发展,如何提高远程教育环境的智能化越来越成为远程教育专家头痛的问题。将数据挖掘技术引入到远程教育领域是对提高远程教育环境智能化的一种方式。本文在研究了经典的关联规则算法 Apriori 之后,提出了类 Spriori 的数据挖掘算法分析学生频繁访问的页面路径,用以提供有用的信息给网络课程设计者以及授课老师,解决学生在网络学习过程中产生的"信息迷航"问题。

1　数据挖掘概述

数据挖掘(Data Mining - DM)简单的说,就是从大量的数据中提取或挖掘知识。就是应用一系列技术从大型数据库或数据仓库中提取人们感兴趣的信息和知识,这些知识或信息是隐含的,事先未知而潜在有用的,提取的知识表示为概念、规则、规律、模式等形式。也可以说,数据挖掘是一类深层次的数据分析[1],数据挖掘应该更正确地命名为"从数据中挖掘知识"。

从不同的视角看,数据挖掘可以分为:

(1) 根据发现知识的种类分类:总结规则挖掘、特征规则挖掘、关联规则挖掘、分类规则挖掘、聚类规则挖掘、趋势分析、偏差分析、模式分析等。

(2) 根据挖掘的数据库分类:关系型、变量型、面向对象型、主动型、空向型、时间型、文本型、多媒体、异质数据库等。

(3) 根据采用的技术分类:人工神经网络、决策树、遗传算法、最临近技术、规则归纳、可视化等。

2　关联规则数据挖掘

关联规则数据挖掘的一个典型例子是购物篮分析。市场分析员要从大量的数据中发现顾客放入其购物篮中的不同商品之间的关系。如果顾客

* 收稿日期:2006-11-17
作者简介:张延萍(1967-),女,上海市人,在读硕士研究生,主要研究方向为数据分析与应用。

买牛奶,他也购买面包的可能性有多大?什么商品组或集合顾客多半会在一次购物时同时购买?例如,买牛奶的顾客有80%也同时买面包,或买铁锤的顾客中有70%的人同时也买铁钉,这就是从购物篮数据中提取的关联规则。

2.1 关联规则

关联规则是如下形式的逻辑蕴涵:$A \Rightarrow B$,其中$A.B$是项集,$A \in I$,$B \in I$,$A \cap B = \Phi$。一般用两个参数描述关联规则的属性。

可信度(Confidence):可信度即是"值得信赖性"。设A,B是项集,对于事务集D,$A \in D$,$B \in D$,$A \cap B = \Phi$,$A \Rightarrow B$的可信度定义为:可信度($A \Rightarrow B$)=包含A和B的元组数/包含A的元组数;可信度表达的就是在出现项集A的事务集D中,项集B也同时出现的概率。如上面的例子中购买牛奶的顾客中有80%也同时购买了面包,即是关联规则:牛奶⇒面包的可信度为80%。

支持度(Support):支持度($A \Rightarrow B$)=包含A和B的元组数/元组总数。支持度描述了A和B这两个项集在所有事务中同时出现的概率。例如在一个商场中,某天共有1 000笔业务,其中有100笔业务同时买了牛奶和面包,则其"牛奶⇒面包"关联规则的支持度为10%。给定一个事务集D,挖掘关联规则问题就是产生支持度和可信度分别大于用户给定的最小支持度和最小可信度的关联规则。

2.2 关联规则数据挖掘的算法

关联规则挖掘是数据挖掘研究的一个重要分支,关联规则是数据挖掘的众多知识类型中最为典型的一种。目前关联规则挖掘问题已经引起了数据库、人工智能、统计学、信息检索、可视化及信息科学等诸多领域的广大学者和研究机构的高度重视,取得了许多研究成果。

2.2.1 Apriori算法

Apriori算法的有效性,在于它利用了一个非常重要的原理,即Apriori性质:如果一个项集是频繁的,则这个项集的任意一个非空子集都是频繁的。该性质属于一种特殊的分类,也称作反单调性。

算法Apriori——发现频繁项目集

(1) L1 = {large 1 - itemsets};

(2) for (k = 2; Lk - 1 ≠ ϕ; k + +) do begin

(3) Ck = apriori - gen(Lk - 1);//新的候选集

(4) for all transactions t ∈ D do begin

(5) Ct = subset(Ck,t);//事务t中包含的候选集

(6) for all candidates c ∈ Ct do

(7) c. count + +;

(8) end

(9) Lk = {c ∈ Ck | c. count(minsup}

(10) end

首先产生频繁1-项集$L1$,然后是频繁2-项集$L2$,直到有某个r值使得Lr为空,这时算法停止。这里在第k次循环中,过程先产生候选k-项集的集合Ck,Ck中的每一个项集是对两个只有一个项不同的属于$Lk-1$的频集做一个($k-2$)-连接来产生的。Ck中的项集是用来产生频集的候选集,最后的频集Lk必须是Ck的一个子集。Ck中的每个元素需在交易数据库中进行验证来决定其是否加入Lk。

2.2.2 Apriori算法的缺陷

(1)算法产生太多虚假(冗余)的规则。当数据库太大或支撑度、信任度阈值太低时产生的规则太多,用户很难人为地对这些规则做出区分、判断。

(2)算法在效率上存在着问题,主要原因为数据库扫描的次数太多,寻找每个k-频繁项目集($k = 1, \cdots, K$)都需要扫描数据库一次,共需要扫描K次。另外,当模式太长时产生的候选项目集也多得让人无法接受。因此当数据库或K太大时,算法的时耗太大或无法完成。做算法可扩展性也不强,难于推广。

3 在远程教育系统中的应用

基于关联规则数据挖掘的远程教育系统,它能够充分利用站点上积累的丰富的信息,更好的服务于远程教学,它与传统的远程教育系统相比,它能为课程的设计者提供用户的浏览模式,重构页面之间的链接,以符合用户的访问习惯;能根据学生自身的情况,提供不同的学习内容或学习进度,做到因材施教;能为老师提供有关学习课件内容的情况,及时调整课程内容的分布,使这符合教学规律;还能根据学生学习课程的掌握程度,向他们提供一些比课程更深或更浅的内容。

3.1 实现对学生学习路径的挖掘

学习路径挖掘模块的输入数据主要是用户与站点的交互数据,一次登录后的一系列满足一定

条件(如:浏览时间≥60s)的访问路径。对学生学习路径的挖掘包括以下步骤:

(1) 数据预处理过程:预处理过程是数据挖掘过程中最关键的一环。处理质量关系到后面挖掘过程和模式分析过程的质量。预处理过程包括 数据转移、数据净化、数据补充等。

(2) 数据挖掘过程:该过程主要是利用一些数据挖掘算法来挖掘出模式、规则等。本文修改 Apiori 算法对学生学习路径进行挖掘产生规则。例如:访问‘执行存储过程’的学生,其中 45% 也访问‘创建存储过程’。

(3) 模式分析过程:在这个阶段主要是利用一些方法和工具对挖掘出来的模式、规则进行分析,找出感兴趣的模式和规则。可以采用 OLAP 技术(例如 数据立方和类 SQL 语言机制来)可视化解释挖掘出来的规则和模式 。

3.2　修改 Apriorir 算法对学生学习路径挖掘产生规则

数据挖掘第一个主要任务就是获得用户在网上的浏览模式,也就是要了解用户在网上的行为。用户的浏览模式也就是用户浏览 Web 网页的方式,因此通过挖掘用户遍历路径来了解用户的浏览模式。

为了方便论述,现约定如下:

(1) 生成的最大向前引用知识点 MFK 称为事务,用 T 表示,也就是一个知识点引用序列,事务 T 中页面的个数称为该引用的长度,一个有 k 个知识点的事务称为一个 k - 引用。

(2) 如果一个 k - 引用序列 $r1,r2,\cdots\cdots,rk$ 满足以下条件,则是频繁 k - 引用:(包含 $r1,r2,\cdots\cdots,rk$ 序列的 MFK 数/所有的 MFK 数) ≥ Supportmin。

(3) Lk 是所有频繁 k - 引用的集合。

(4) Ck 表示候选集,频繁 k - 引用是满足一定支持度的 Ck。

(5) 最大频繁引用是指不被任何其它频繁引用包含的频繁引用,例如:$L2 = \{AB,BG,AD,CG,CH,BG\}$,$L3 = \{ABE,CGH\}$,那么最大引用序列是 $\{AB,BG,ABE,CGH\}$,AB 等由于被 ABE 包含,因此不是最大频繁引用。

修改过的 Apriori 算法具体描述如下:

(1) L1 = {c ∈ C1 | c. count ≥ minsupport};
(2) C2 = getFullArray();
(3) L2 = {c ∈ C2 | c. count ≥ minsupport};
(4) For(k = 3;LK - 1 ≠ ϕ;k + +)
(5) Ck = gen_candidate(Lk - 1);
(6) For all transaction t ∈ D
(7) Ct = count_support(Ck,t)
(8) For all candidates c ∈ Ct
(9) c. count = c. count + 1;
(10) Next
(11) LK = {c ∈ Ck | C. count ≥ min_sup};
(12) Next
(13) Resultset = resultset ∪ Lk;
(14) Next

其中 D 表示事务数据库,minsupport 表示给定的最小支持度,resultset 表示所有的频繁引用集。

上面的算法和 Apriori 算法表面上几乎一模一样,但是求候选引用集的函数 gen_candidate 不一样。在关联规则中,只要两个 $k-1$ 维最大项集与 $k-2$ 各元素相同就可以合并成一个 K 维候选项;但在挖掘频繁路径时,引用中的页面是有序的,因此不能简单地只要 $k-2$ 个元素相同就行了,需要做如下的修改:$Lk-1$ 中任意两个不同的 $(k-1)$ - 引用中一个去掉第一个元素,另一个去掉最后一个元素后完全相等,则这两个 $(k-1)$ 引用可以合并成一个 k - 引用。

3.3　挖掘结果的运用

学习路径的挖掘结果以文字的方式提供给需要的用户。挖掘结果将根据对大多数学生的学习路径的挖掘,可以提供给网络课程设计者对其设计的网络学习课件进行优化处理,调整各页面之间的链接,使其更加符合学生学习的习惯。学生学习路径的挖掘结果可以反馈给远程教育系统,由系统动态生成某些链接,满足学生的特定需要,实现对学生的个性化教学。对某一个特定的学生学习路径的挖掘可以使教师掌握该学生对于某门课程的学习状况和学习进度,可以对学生进行学习建议。挖掘的结果也可以使教师了解学生学习课程的掌握程度,以此为根据提供学生一些比已有课程更叫深入或者更加浅显的教学内容。挖掘的结果也可提供给学生本人,这样可以使学生对于自己的学习状况有一个全面地了解,有助于学生进行自主学习。

(下转第 53 页)

参考文献:

[1] 雨阳隆春. 深入 JSP 网络编程[M]. 北京:清华大学出版社,2001.
[2] 姜晓铭,陈武,等. JSP 程序设计与实例分析教程[M]. 北京:清华大学出版社,2001.
[3] 飞思科技产品研发. JSP 应用开发详解(第二版)[M]. 北京:电子工业出版社,2004.
[4] 郑阿奇. Oracle 实用教程[M]. 北京:电子工业出版社,2004.
[5] 东方人华. Oracle 9i[M]. 北京:清华大学出版社,2003.

Online Teaching Management System Based on Jsp、WebLogic and Oracle

WANG Ming – hui
(The Modern Educational Technology Center of Yancheng Institute of Technology, Jiangsu Yancheng 224003, China)

Abstract: This teaching management system is an interactive application system, which is based on B/S structure and adopts Oracle as its background database and WebLogic plus Jsp as its running environment. It is accomplished with display module, information management module and teacher management module. The application server of this system is Weblogic, but the main characteristic of Jsp is cross – platform, so it is used in other application server.

Keywords: Jsp; Weblogic; Oracle; modulization

(上接第 49 页)

4　结论

本文在研究了经典的关联规则算法 Apriori 之后,提出了类 Spriori 的数据挖掘算法分析学生频繁访问的页面路径,用以提供有用的信息给网络课程设计者以及授课老师,来解决学生在网络学习过程中产生的“信息迷航”问题。随着数据挖掘技术的发展,相信远程教学系统中将有更多方面将应用到数据挖掘技术。

参考文献:

[1] 陈怀东,张小真. 一种基于多 Agent 的网上协作自适应学习模型 MASAM 研究[J]. 现代教育技术研究与应用,2002(2):22 – 24.
[2] Chad Creighton, Samir Hanash, mining gene expression databases for association rules[J]. Bioinformatics, 2003,19(1): 79 – 86.

Study and Application of Data Mining Based on Association Rules

ZHANG Yan – pin
(The Office of the President, Yancheng Institute of Technology, Jiangsu Yancheng 224003, China)

Abstract: After the of study of the apriori algorithm in classical association rules, a date mining algorithm like spriori is proposed to analyse the pages frequently accessed by students, as it could provide useful information for the network course designers and teachers to solve the information confusion problems among students in network learning.

Keywords: association rules; Apriori; data mining; distance instruction