《电子技术应用》 www.ChinaAET.com

基于负关联规则的 Web 使用挖掘技术及发展趋势*

杨 斌,董祥军
(山东轻工业学院 信息科学与技术学院,山东 济南 250353)

摘 要:介绍了 Web 使用挖掘各阶段的主要工作以及相关技术,重点介绍模式发现阶段负关联规则的应用,并对将来 Web 使用挖掘领域的研究作了展望。

关键词:数据挖掘;Web 使用挖掘;负关联规则
中图分类号:TP311 **文献标识码**:A

Technology and development tendency of Web usage mining based on negative association rules

YANG Bin, DONG Xiang Jun
(School of Information Science and Technology, Shandong Institute of Light Industry, Jinan 250353, China)

Abstract: The main work and related technology about various stages of Web usage mining are presented, the paper expatiates the application of the negative association rules in the stage of pattern discovery. Some future works on research field of Web usage mining are also presented.

Key words: data mining; Web usage mining; negative association rules

近年来,Internet 的普及和发展改变了人们获取信息的方式,人们通过 Web 接触到了大量的数据和信息,但由于 Web 上的信息是半结构化的、动态的,不容易发现数据中存在的关系和规则,也无法根据现有的数据预测未来的发展趋势。Web 数据挖掘[1]就是将传统的数据挖掘技术与 Web 结合起来进行数据挖掘。随着 Web 技术的进一步发展,Web 站点设计、个性化服务、电子商务等各项工作也变得愈加复杂,而 Web 数据挖掘成了解决上述问题的有效的方法。Web 挖掘分为 3 类:Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘。

Web 使用挖掘是从服务器端记录用户访问日志或从用户的浏览信息中抽取感兴趣的模式,通过分析这些数据可以帮助理解用户的行为,对提供个性化服务与定制、改进 Web 系统性能和结构、改善 Web 站点结构、为商业组织提供商业智能和向用户推荐页面等方面都有重要的理论和实际意义。Web 使用挖掘也因此成为目前国内外研究的一个热点。

1 负关联规则的相关理论及研究现状

1.1 负关联规则的相关理论

传统的关联规则 AR(Association Rule) 是 $A \Rightarrow B$ 的形式,用于挖掘顾客事务数据库中项集间的关联关系,最初由 AGRAWAL R 等人于 1993 年首先提出,并于 1994 提出了一种快速算法[2]。这些算法仅能用来发现强模式,即那些具有高频率和强相关的显式模式,但是数据库中还存在许多采用这些挖掘技术所不能发现的隐式模式,而其中之一便是负关联规则,它具有低频率、强相关的性质,表现了数据项目集间的不易直接觉察到的强相关性质。当决策者想知道“某些有利因素出现时,哪些不利因素很少出现”的时候,负关联规则就变得非常重要。

负关联规则的支持度和置信度的定义为:设 $I=\{i_1,i_2,\cdots,i_m\}$ 是项的集合,D 是数据库事务的集合,其中每一个事务 T 是 I 中一组项目的集合,即 $T \subseteq I$。一个负关联规则是形如 $A \Rightarrow \neg B$、(或 $\neg A \Rightarrow B$、$\neg A \Rightarrow \neg B$) 的蕴涵式,

* 基金项目:山东省自然科学基金(Y2007G25),山东省优秀中青年科学家奖励基金项目(2006BS01017)

《电子技术应用》 www.ChinaAET.com

$A \subseteq I, Y \subseteq I$,而且,$A \cap B=\phi$。为了方便只定义 $A \Rightarrow \neg B$ 的支持度和置信度,则,规则 $A \Rightarrow \neg B$ 在 D 中的支持度(support,简记为 S)是事务集中包含 A 和 $\neg B$ 的事务数与所有事务数之比,记为 $S(A \cup \neg B)$;规则 $A \Rightarrow \neg B$ 在事务数据库 D 中的置信度(confidence,简记为 C)是指包含 A 和 $\neg B$ 的事务数与包含 A 的事务数之比,记为 $C(A \cup \neg B)$。

负关联规则挖掘就是在数据库 D 中筛选出所有满足用户指定的最小支持度 $minsupp$ 和最小置信度 $minconf$ 的负关联规则 $A \Rightarrow \neg B$(或 $\neg A \Rightarrow B$、$\neg A \Rightarrow \neg B$),$A$、$B$ 分别为频繁项集。

1.2 负关联规则的研究现状

WU Xin Dong 等提出了 $A \Rightarrow \neg B$ 等 3 种形式的负关联规则 NAR (negative AR),而 $A \Rightarrow B$ 型的关联规则相应地称为正关联规则 PAR (Positive AR)[3],并且给出了一个 PR 模型以及能够同时挖掘正关联规则和负关联规则的算法[4],该算法以传统的 Apriori 来挖掘频繁项集和非频繁项集,在挖掘频繁项集中正关联规则的同时,能够清楚地挖掘非频繁项集中的 $A \Rightarrow \neg B$、$\neg A \Rightarrow B$ 以及 $\neg A \Rightarrow \neg B$ 型负关联规则。

Dong 等人利用了支持度-置信度框架,提出了一种 PNARC 模型[5],根据已知的正关联规则的支持度和置信度,计算负关联规则的支持度和置信度:设 $A, B \in I$, $A \cap B=\phi$,则有:

(1) $supp(\neg A)=1-supp(A)$;

(2) $supp(A \cup \neg B)=supp(A)-supp(A \cup B)$;

(3) $supp(\neg A \cup B)=supp(B)-supp(A \cup B)$;

(4) $supp(\neg A \cup \neg B)=1-supp(A)-supp(B)+supp(A \cup B)$;

(5) $conf(A \Rightarrow \neg B)=\frac{supp(A)-supp(A \cup B)}{supp(A)}=1-conf(A \Rightarrow B)$;

并采用相关性检验方法,能够有效挖掘出频繁项集中的正、负关联规则,检测并删除相互矛盾的规则;随后又给出一种基于多置信度和 χ^2 检验的挖掘正负关联规则的方法,进而提出了一种 PNARMC 算法[6],该算法在正确产生正负关联规则的基础上,可以灵活地控制关联规则的数量。

2 Web 使用挖掘的过程及发展现状

2.1 Web 使用挖掘的过程

国外最早研究 Web 使用挖掘是从 1996 年开始的,CHEN MS. 等在参考文献[7]中首先提出了将数据挖掘的技术应用于 Web 领域,从而发现 Web 中隐藏的知识。Web 使用挖掘过程通常可以分为 3 个步骤:数据收集与预处理、模式发现、模式分析,其结构如图 1 所示。

2.1.1 数据收集与预处理

Web 使用挖掘的数据来源于服务器端、客户端和代理服务器端。数据获取就是从服务器日志、引用日志、代理服务器日志和客户端收集这些数据。一般情况下,收集到的数据中可能有冗余的和遗漏的数据,因而要对数据进行预处理以得到适合于模式发现的数据格式,处理的过程一般包括数据清洗、用户识别、会话识别、事务识别和路径补充等。

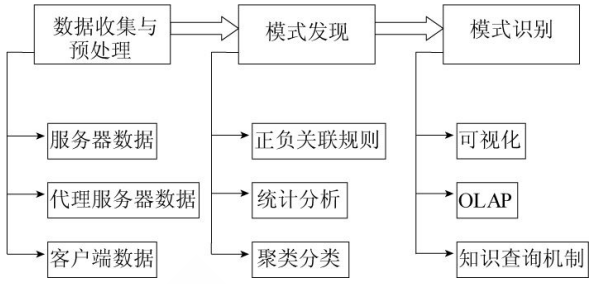

图 1 Web 使用挖掘的过程

2.1.2 模式发现

模式发现是 Web 挖掘的关键部分,模式发现就是利用挖掘算法挖掘出有效的、新颖的、潜在的、有用的及最终可以理解的信息和知识,它集合了数据挖掘、机器学习、统计学和模式识别等研究领域的算法和技术。目前常用的模式发现技术包括关联规则、路径分析、聚类等。

关联规则挖掘就是挖掘出用户在一个会话期间在服务器上访问的页面和文件之间的关系,找出经常一起出现的相关页面,也就是找出支持度和置信度满足指定阈值的关联规则。在实际的 Web 站点设计中往往根据挖掘出来的关联规则,合理地组织网站的框架和连接结构,方便用户访问,使用户长久保持对访问站点的兴趣,有助于提高网页的点击率,更好地增加网站的经济价值。

2.1.3 模式分析

模式分析是 Web 使用挖掘的第 3 阶段,这个阶段是从模式发现过程的结果中去除不相关的规则或模式以及抽取有兴趣的规则或模式。目前研究所用的模式分析方法和工具包括类 SQL 查询机制、OLAP 和可视化等。

2.2 基于负关联规则的 Web 使用挖掘的发展现状

Web 使用挖掘近年成为一个热点,参考文献[8]介绍了 Web 使用挖掘的理论背景、基本原理以及 Web 使用挖掘在挖掘方法和应用方面的研究成果。参考文献[9]分别对 Web 使用挖掘的 3 个过程即数据收集与预处理、模式发现、模式分析分别进行了难点分析,并对未来的研究重点进行了展望。

近年来,有一部分学者研究了基于正关联规则的 Web 使用挖掘,并且挖掘出很多实用的、有价值的信息。参考文献[10]提出了在 Web 使用数据中挖掘有时间限制的关联规则,描述了如何把该方法用于购物篮分析。参考文献[11]设计了一个基于 Web 日志的个性化推荐系统,任务的核心是建立一个用户特征文件,用户的信息是在 Web 日志中学习来的,以此来构建用户特征文件。现有的基于关联规则的 Web 使用模式挖掘,利用 Apriori 算法找出频繁项集,再挖掘出正关联规则,不能

《电子技术应用》 www.ChinaAET.com

发现负关联规则,而且使用 Apriori 算法,多次地扫描数据库并产生庞大的侯选集,这些操作都需要消耗大量的时间和内存空间,在向用户推荐时,可能含有用户不需要的页面, 在挖掘关联规则方面也只是发现正关联规则,挖掘出的关联规则不全面。

3 Web 使用挖掘中负关联规则的发展趋势

Web 使用挖掘是数据挖掘在 Web 世界的延伸,作为一个新的研究领域正在蓬勃发展,它很大程度上改善了人们工作和生活的方式, 把负关联规则应用到 Web 使用挖掘中是一个新的研究方向,还有很多工作要做。为此, 从以下几个方面对基于负关联规则的 Web 使用挖掘进行了展望。

(1) 负关联规则应用到 Web 使用挖掘中可以发现感兴趣的规则和模式,正关联规则挖掘在支持度置信度框架下能发现页面之间比较紧密的相互关系,但它不能发现 Web 网页之间的互斥关系, 反映到 Web 页面中就是用户点击 A 页面能够减少访问 B 页面的可能性,而这种关系可能在网站布局中是非常重要的。负关联规则挖掘作为一种隐式的挖掘方法具有低频率、强相关的规则,在实际应用中将为网站管理员提供更有价值的信息或能更好地满足用户个性化的需要。

(2) 现有的负关联规则算法十分丰富,而且各有所长,各种算法的变形也是层出不穷,在具体的 Web 环境下找出一个有效的、复杂度低的算法还有待于研究,且已有的频繁路径发现方法过于复杂和需要人工计算,使用经典算法 Apriori 有重复扫描数据库和产生侯选集需要消耗大量的时间和内存空间的缺陷。因此,如何编写更高效的匹配 Web 环境的负关联算法、自动地发现用户频繁访问路径还有待研究。

(3) 目前, 基于负关联规则的 Web 使用挖掘方法的研究比较少, 可以在 Web 环境中把正负关联规则都挖掘出来,正负关联规则可以形成互补,从而能更好地为网站及企业管理者提供决策支持,因此基于负关联规则的 Web 使用挖掘是一个新的研究方向。

参考文献

[1] COOLEY R, MOBASHER B, SRIVASTAVA J. Web mining:Information and pattern discovery on the World Wide Web [C]. In:proc.of the 9th Int'l Conf. on Tools With Artificial Intelligence (ICTAI'97), CA, 1997:558-567.

[2] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules in large database [C]. Proceedings of the 1994 International Conference on VLDB. San Francisco: Morgan Kaufmann Publishers, 1994: 487-499.

[3] WU Xin Dong, ZHANG Cheng Qi, ZHANG Shi Chao. Mining both positive and negative association rules [C]. Proceedings of the 19th International Conference on Machine Learning (ICML-2002). San Francisco:Morgan Kaufmann Publishers, 2002: 658-665.

[4] WU X, ZHANG C, ZHANG S. Efficient mining of both positive and negative association rules[J]. ACM Transactions on Information Systems, 2004, 22: 381-405.

[5] DONG X,WANG S,SONG H.Study of negative association rules [J]. Beijing Institute of Technology Journal, 2004,24 (11): 978-981.

[6] DONG X, SUN F, HAN X, et al. Study of positive and negative association rules based on multi-confidence and Chi-Squared test [J]. LNAI 4093, Springer-Verlag Berlin Heidelberg, 2006: 100-109.

[7] CHEN M S, PARK J S, YU P S. Efficient data mining for path traversal patterns in a web environment [J].IEEE Trans on Knowledge and Data Eng,1998,10(2):385-390.

[8] 黄浩, 王建军. WEB 使用挖掘研究 [J]. 计算机系统应用, 2008 (1).

[9] 朱志国,邓贵仕. Web 使用挖掘技术的分析与研究[J].计算机应用研究, 2008 25(1): 29-36.

[10] HUYSMANS J, MUES, CHRISTOPHE, et al. Web usage mining with time constrained association rules [C]. ICEIS 2004-Proceedings of the Sixth International Conference on Enterprise Information Systems, 2004: 343-348.

[11] SUTHEERA P, TSUJI H. Mining web logs for a personalized recommender system [C]. ITRE 2005-3rd International Conference on Information Technology: Research and Education - Proceedings, 2005: 445-448.

(收稿日期:2008-12-18)

智能领先 极速酷睿™

--2009 年英特尔寒促正式启动

(2009 年 12 月 21 日-北京)随着寒假和元旦的临近,英特尔"智能领先,极速酷睿" 2009 寒促也于日前正式拉开序幕。此次寒促活动覆盖了全国 108 个城市,不仅将举办各种丰富多彩的线下和线上活动,更有"新三剑侠"和"大小乔"等台式机和笔记本的明星处理器轮番上阵,势必给消费者带来新一轮的惊喜,在寒冬为广大用户奉上一场科技盛宴。

"此次寒促推出新三剑侠中, 包括了具有智能特性的英特尔酷睿 i5 处理器, 将为今冬的 PC 市场燃起智能电脑普及的一把火,"英特尔中国区市场与渠道部总经理张文翊女士表示:"英特尔 2009 暑促三剑侠产品自推出以来,受到消费者的追捧,已经逐渐成为市场主流。寒促新三剑侠以智能性能为特点,更清晰地定义了市场上的三种主流需求,消费者将直接体验到英特尔产品在设计、游戏和数字生活等领域的全新应用。"

(英特尔公司供稿)

《电子技术应用》 www.ChinaAET.com