基于关联规则的 Web 运行数据挖掘技术分析

岳 千

（西安工业大学 教务处，陕西 西安 710021）

摘 要：在数据库信息技术水平不断提高的背景下，数据挖掘技术也得到了广泛应用，能够从海量信息中寻找有价值的信息。Web 系统为国内外普遍应用的信息发布渠道，其商业价值显著。在 Web 站点上实现运行数据挖掘，可以让信息技术更好地服务于人们的生产与生活。笔者结合数据库信息挖掘技术的发展现状，对基于关联规则的 Web 运行数据挖掘技术进行解析，以供各位同仁参考。

关键词：关联规则；Web 运行数据挖掘技术；Apriori 算法；UFAPA 算法

中图分类号：TP311.13 **文献标识码：**A **文章编号：**1003-9767（2022）03-207-03

Analysis of Web Operation Data Mining Technology Based on Association Rules

YUE Qian

(Academic Affairs Office, Xi'an Technological University, Xi'an Shaanxi 710021, China)

Abstract: Under the background of the continuous improvement of database information technology, data mining technology has also been widely used, which can find valuable information from massive information. The Web system is a widely used information release channel at home and abroad, and its commercial value is significant. Implementing data mining on Web sites can make information technology better serve people's production and life. In combination with the development status of database information mining technology, the author analyzes the Web operation data mining technology based on association rules for the reference of colleagues.

Keywords: association rules; Web operation data mining technology; Apriori algorithm; UFAPA algorithm

0 引言

为满足人们工作、学习、生活的需求，各种网站如雨后春笋一般涌现，海量的信息给用户带来了许多便利，促使用户的视野日渐开阔[1]。同时，内容繁多的网络世界给用户获取目的信息造成了一定困难，若盲目寻找，则如大海捞针一样耗时耗力。此外，人们日常生产生活和信息化技术之间的关联越发密切，通过大数据技术进行数据挖掘可以为企业决策提供参考[2]。近年来，Web 运行数据挖掘技术是颇受关注的信息化技术之一。Web 系统是现阶段 Internet 信息发布的主要渠道，其应用潜力及商业价值显著[3]。根据该系统运行情况来看，Web 数据本身的特性导致数据挖掘工作十分复杂，且这一技术和基于数据库的数据挖掘有所不同，其过程更为复杂、严谨。因此，在关联规则基础上探索 Web 运行数据挖掘技术。

1 Web 运行数据挖掘的定义及分类

1.1 定义

Web 数据挖掘是在 Web 环境下应用数据挖掘技术，是对信息科学技术、数据挖掘技术以及 Web 技术等众多领域先进技术的综合应用，可以从海量的 Web 文档集合、站点内浏览数据内挖掘出有价值的信息[4]。换言之，Web 数据挖掘是在对大量数据分析的基础上进行归纳性推理，对客户行为进行预测，帮助企业决策人员对市场策略进行调整，辅助决策人员作出正确决策[5]。

1.2 分类

Web 数据挖掘包括内容挖掘、结构挖掘、日志挖掘 3 大部分。

作者简介：岳千（1990—），女，天津人，硕士研究生，助教。研究方向：数据挖掘。

1.2.1 Web 内容挖掘

Web 内容挖掘是对 Web 文档内容和描述的文档信息规律及模式进行挖掘。因 Web 数据自身结构特点，要想从数据源中实现数据自动化挖掘，面临一定的挑战。现阶段，Web 内容挖掘属于重要研究方向，是基于用户角度将数据中的有价值信息挖掘出来，为用户决策提供参考。

1.2.2 Web 结构挖掘

Web 站点中的页面链接关系与组织结构内往往存在大量信息，而 Web 结构挖掘即是从其中挖掘出有价值的模式或知识。这些模式或知识可在 Web 站点应用，提高站点的检索速度，让用户获得更优良的搜索体验。

1.2.3 Web 日志挖掘

Web 日志挖掘即 Web 使用挖掘，是从日志文件中挖掘所需信息，以探索用户访问的规律或模式。日志可对用户登录网站的地址、时间、浏览器、页面路径、请求方法以及服务器端信息等进行记录，以此反应用户群体的共性与个性。

Web 运行数据挖掘技术是 Web 内容挖掘、结构挖掘及日志挖掘的整合，其涵盖点更为全面，分析 Web 运行数据挖掘技术有助于提升客户检索信息的效率 [6]。

2 关联规则定义

在数据库技术不断发展的背景下，信息规模越发庞大。从表面看并无任何联系的数据，其内部存在巨大商业价值。例如通过商业平台背后的交易数据，可以发现一些关联知识，协助企业管理人员作出商业决策，实现交叉营销。而在这一过程中，则会用到关联规则 [7]。

关联规则指的是在大量数据内对数据项间规则或联系进行挖掘的规则，可对独立事务彼此存在的关联性或依赖性予以反映。若单一事务和多个事务之间存在关联，则可通过已知事务对另一个事务进行预测。换言之，关联规则将事务作为研究对象，通过各种数据挖掘方法及技术对数据项的内部关系进行挖掘，分析事务 X 的出现对事务 Y 出现的影响。

目前，数据库信息技术研究人员十分重视数据库项目集合内的关联规则算法，并且展开了相关研究。在应用关联规则算法时，可引入并行思维模式，提升数据挖掘技术处理速度及效率。搜索算法是在对数据集内项目信息统一读取后展开计算，可以通过一次扫描查找出所有频繁项目集。值得注意的是，在关联规则的算法数据挖掘之中，需要先确定运用哪种关联规则来实现数据处理，然后才能开展后续的计算工作。

3 基于关联规则的 Web 运行数据挖掘过程

Web 运行数据挖掘技术种类很多，应用最广泛的是借助多维数据立方体来进行多维关联规则挖掘、转换后挖掘以及直接挖掘。多维关联规则挖掘将有关日志文件形成关系数据库，根据数据库来构建多维 Web 数据立方体。转换后挖掘是将 Web 日志文件中的属性经过转换形成摩尔矩阵，若匹配则其属性值视作 1，否则视作 0。直接挖掘是直接挖掘日志文档，不用特殊转换。

上述常用的 3 种方法均未充分考虑 Web 日志特点。Web 日志文件的内容复杂，文档结构具有半结构化特点，且在不断更新 [8]。笔者认为，可以先对 Web 运行数据进行对应处理，做好数据预处理后将其存储在关系型数据库内，之后再实现数据挖掘。

3.1 数据预处理

Web 运行数据挖掘中的数据格式与数据挖掘需求并不完全相符，故而需要采取中间操作，如清洗、转换等，将不合适数据格式转换为其他数据格式，并在数据库内保存。Web 运行数据的预处理方法及技术并非唯一，需要结合不同应用场景和挖掘目的采取不同方法进行预处理。只有目的明确，挖掘才更有效率 [9]。数据预处理包括数据源收集、净化、识别用户、识别用户会话、页面过滤以及用户路径补充等。

3.2 模式挖掘与分析

模式挖掘又称模式发现，是数据挖掘的核心，需通过人工智能等多种算法来实现。完成挖掘后，需要对挖掘结果进行模式分析。挖掘获取的模式并非全都有价值，也并非全部正确。借助相关分析工具、科学技术、科学方法，将模式转变成有价值含量高的知识，之后运用图形界面将这些知识传递给决策人员，由决策人员根据这些知识来作出决策 [10]。

4 基于关联规则的 WeB 运行数据挖掘技术

4.1 数据分析及清理

4.1.1 连续属性离散化

在数据挖掘过程中，连续属性的离散化是重要的问题之一。数据挖掘算法需要预先离散化连续属性数据，而关联知识挖掘需减少给定连续属性值的个数。离散化任务指的是将连续属性取值区间、取值范围划分成若干个区间，各区间与一离散化符号对应，即用区间标号来取代实际数据值。

以时间作为考察属性，在离散化过程中将 5:00—8:00 定义为早上、8:00—12:00 定义为上午、12:00—14:00 定义

为中午、14:00—18:00 定义为下午、18:00—22:00 定义为晚上，即区间划分为 $[a,a+(b-a)/N],[a+(b-a)/N,a+2(b-a)/N,],\cdots,[a+(N-1)*(b-a)/N,b]$，其中 N 为用户设定的离散值个数，a 与 b 为各区间的时间点。

4.1.2 预处理

数据的预处理过程重点在于数据过滤及补充，去掉重复的记录。这一阶段的主要任务是从原始文件中选取用户浏览模式，发现算法可用的规范化数据。这一工作结果将会对算法处理结果准确性、可信度造成影响。其中，数据过滤就是将不需要的数据删除，提取用户 IP 地址、请求页面等，获取用户浏览过的页面。用户识别是将用户与请求页面彼此关联的重要过程，主要是用户能否访问站点。在这一过程中，不仅会用到服务器的日志，而且还需借助站点拓扑结构来实现。

4.1.3 用户识别及路径补充

基于本地缓存、防火墙及代理服务器的存在，对每一个用户进行识别变得十分复杂。Web 运行数据挖掘中，部分可以通过启发式的规则将用户识别出来。在用户识别中，还需要确定访问日志中有没有用户的重要请求未被记录的情况。若日志文件中并未记录，且这一数据能对用户行为进行补充，则需通过路径补充来实现。

4.2 具体算法

基于关联规则的 Web 运行数据挖掘技术中，Apriori 算法是最经典的算法之一，其在关联规则分层法中属于分层算法，数据集扫描的次数和最大频繁项目集数保持相等。在挖掘布尔关联规则中，Apriori 算法的访问路径是在 Web 日志中，通常寻找支持度比预定值大的访问路径。

与 Apriori 算法不同，UFAPA 算法中的页面是无序的，考虑所有信息综合之后产生的效应，不能对个体行为予以准确反映。在 Web 大型数据库中，调用数据时会用到海量的系统资源，所需系统时间较长，极易死机，导致挖掘工作中断，影响最终工作结果。为提升算法的效率，在确保候选频繁路径不丢失的前提下，减少数据可扫描的次数，同时尽可能减少候选频繁路径的数量。若两次遍历之后采用一种算法计算，可促使长度 k 为 2 以上的所有频繁路径集均产生，这时只要再对数据遍历一次就可获得所有频繁路径集，这是 UFAPA 算法的实现思路。根据这一算法思想，如果 BP 为频繁路径，则 BP 所有的子路径 BP’也为频繁路径。按照这一算法挖掘的过程，可以分成长度是 0 的频繁遍历路径，之后再次挖掘可以再次获得频繁遍历路径。在这一背景下，就可将长度是 k 的频繁遍历路径挖掘问题转换成长度是 0 的频繁遍历路径或长度是 1 的频繁遍历路径，这样成本会更小。

5 结 语

数据挖掘改变了传统数据低层次查询的模式，从数据内知识挖掘的角度来为决策提供所需的信息。通过关联规则可以发现不同数据集之间的联系，在此基础上实现 Web 运行数据挖掘，能够让 Web 管理员实现对 Web 站点有效管理，为用户快速、准确获取所需信息提供支持。Web 日常数据源与挖掘模式存在特殊性，未来在 Web 运行数据挖掘方面仍需不断探索和创新。

参考文献

[1] 程军锋 .Web 数据挖掘研究 [J]. 重庆三峡学院学报 ,2021(3):43-45.

[2] 刘莉瑶，鲍正德，李晨曦 . 浅谈 Web 数据挖掘技术在信息管理中的实际应用 [J]. 计算机系统网络和电信 ,2019(1):19-22.

[3] 孙红，李存进 . 融合遗传算法和关联规则的数据挖掘方法改进 [J]. 数据采集与处理 ,2019,34(5):863-871.

[4] 王琼，杨明杰，闫润珍 . 大规模数据库中关联规则的发现与研究 [J]. 科技经济导刊 ,2020(34):47-48.

[5] 张宏，吕悦晶 . 基于大数据挖掘技术的车载自组织网络状态异常检测 [J]. 汽车技术 ,2019(10):48-52.

[6] 郝林倩 . 基于关联规则的数据挖掘算法分析 [J]. 太原学院学报 (自然科学版),2020,38(3):42-45.

[7] 朱勇，沈士强 . 基于关联规则挖掘的 Oracle 数据库审计分析系统的设计 [J]. 数字技术与应用 ,2019(4):56-57.

[8] 曾子贤，巩青歌，张俊 . 改进的关联规则挖掘算法 :MIFP-Apriori 算法 [J]. 科学技术与工程 ,2019,19(16):216-220.

[9] 王庆桦 . 基于数据挖掘技术的图书馆个性化快速推荐算法研究 [J]. 现代电子技术 ,2019,42(5):149-151.

[10] 杨井荣，侯向宁 . 正负关联规则数据挖掘算法研究 [J]. 计算机技术与发展 ,2020,30(11):64-68.