基于FP-GROWTH算法的关联规则挖掘算法研究

陈 寅
（南京华苏科技有限公司，江苏 南京 210012）

摘 要：互联网世界的数据每年都在成倍增长，但是对用户有用的信息却好像在减少，用户淹没在数据的海洋中，虽然类似于Google这样的搜索引擎可以帮用户找到需要的信息，但是正确率和查全率都不尽如人意。数据挖掘是兴起于20世纪90年代的一项用于决策支持的新技术。FP-GROWTH算法只进行2次数据库扫描。它不使用候选集，直接压缩数据库成一个频繁模式树，最后通过这棵树生成关联规则。文章研究FP-GROWTH算法理论的同时实现了一个简单算法演示的系统。系统包括算法的执行，对数据库的修改、查询、删除的操作。最后，对FP-GROWTH算法和Apriori算法进行了比较。

关键词：数据挖掘；关联规则；FP-GROWTH算法；候选集；频繁模式树

1 基于FP-GROWTH算法的关联规则挖掘算法

1.1 FP-GROWTH算法的基本思想

FP-GROWTH算法采用归纳分散的策略，对数据库进行第一次扫描，把数据库中的频繁项集压缩到一棵频繁模式树（FP-Tree），同时依然保留其中的关联信息，随后再将FP-Tree分化成一些条件数据库，每个条件数据关联一个频繁项，然后再分别对这些条件库进行挖掘。FP-GROWTH算法核心思想如下所示：输入事务数据库D；最小支持度阈值min_sup。输出频繁模式的完全集。FP-tree的产生可由下例进行简单的介绍[1-3]。

我们给出了一个简单的数据集合{1，3，4}，{2，4，5}、{2，4，6}。先对数据库进行一次扫描，根据集合中项的出现频率可以得出一个数据集合{4，2，1，3，5，6}（项的次序按出现频率由高到低排列），可以把这个集合认为是对数据库扫描后进行了一次整理，生成了一个新的数据库。由这个集合按项出现的频率生成FP-Tree。我们先读取第一个集合，并按集合中项的出现频率决定是否优先插入，插入后该节点的计数加1，同样的方法再插入第二个集合，如果集合中项与FP-Tree中已有的节点重复，那么该节点计数加1，如果不重复，插入该项并且该节点计数加1，重复上述操作直至完成所有项的插入。具体实现步骤如图1所示[4-7]。

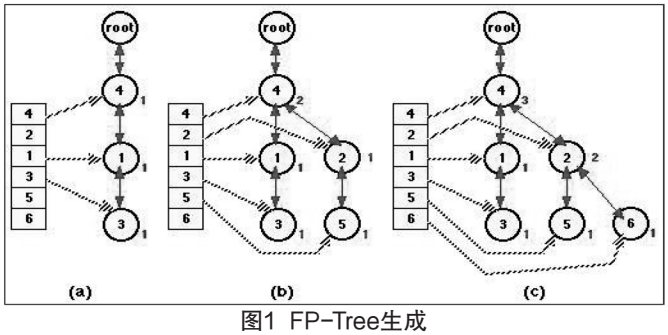

图1 FP-Tree生成

1.2 基于FP-GROWTH算法的关联规则挖掘算法的系统实现

1.2.1 算法实现环境

从上面的阐述中，可以了解到系统的实现采用了C++和SQL server，在这里我们对编译环境进行简介，并说明它们的优点。

Visual C++是一个功能强大的可视化软件开发工具。自1993年Microsoft公司推出Visual C++1.0后，随着其新版本的不断问世，Visual C++已成为专业程序员进行软件开发的首选工具[8]。

SQL Server 2000是为迅速提供可伸缩性电子商务、企业及数据仓库解决方案而开发的完整数据库与分析软件产品。SQL SERVER 2000定位于Internet背景下的数据库应用，它为用户的Web应用提供了一款完善的数据管理和数据分析解决方案。同时SQL SERVER 2000还是Windows分布式网络架构（Distributed Internet Architecture，DNA）架构的一个核心组件。它极大地缩短了用户开发电子商务、数据仓库应用的时间。SQL SERVER 2000还提供对扩展标示语言支持（Extensible Markup Language，XML）和HTTP的全方位支持[9-10]。

1.2.2 系统功能设计

系统功能主要包括以下几点：（1）对数据库有基本的操作功能，如查询、修改、删除等操作。（2）采用C++来实现FP-GROWTH算法。（3）界面可视化[11-15]。

1.2.3 数据库设计

首先我们需要挖掘商品之间的潜在关联关系，那么必须知道商品的最基本的一些属性，建立一个名叫shangpin的表，其中ID为主键，数据类型为varchar，长度为50.其中还包括了Name（商品名称）；Pdate（生产日期）；Sdate（销售时间）；Price（商品价格）[16]。具体设置如表1所示。

表1 shangpin

名称	数据类型	大小	索引
ID	varchar	50	主键
Name	varchar	50	
Pdate	datetime	8	
Sdate	datetime	8	
Price	float	8	

在执行算法之后，人们真正感兴趣的是顾客所购买的商

作者简介：陈寅（1986— ），男，江苏扬州人，高级工程师，学士；研究方向：大数据挖掘。

品之间潜在的关联规则，再建立一张名叫sale的表，在该表内所有的购买项目生成了联合主键item1代表了顾客购买的一件商品，以此类推item2，item3分别为购买的第二、三件商品，所有的项组成了顾客一次购买的商品所组成的集合。

表2 sale

名称	数据类型	大小
Item1	varchar	50
Item2	varchar	50
Item3	varchar	50
Item4	varchar	50
Item5	varchar	50
Item6	varchar	50
……	varchar	50

1.2.4 系统界面设计

系统的界面，主要提供了数据库的基本操作界面，FP-GROWTH算法的执行界面，在进入程序后会弹出用户登入界面，用户登入后，为方便用户对系统进行针对性的操作，笔者设计了工具选择窗口。具体设计如图2所示。

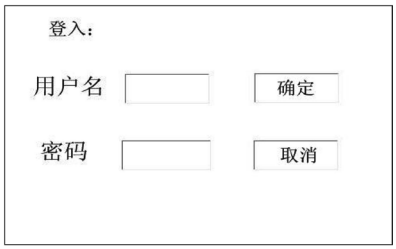

图2 登入设计界面

登入界面包含了用户名和密码，可对用户输入的用户名、密码进行验证。如果错误提示重新输入，输入错误次数超过3次时自动退出系统。

在数据库的界面上人们能够实现数据的查询、添加、修改、删除等操作。界面中还包含了商品序列号、商品名称、查询方式等操作窗口（见图3）。这些使得人们更能直观地进行使用。

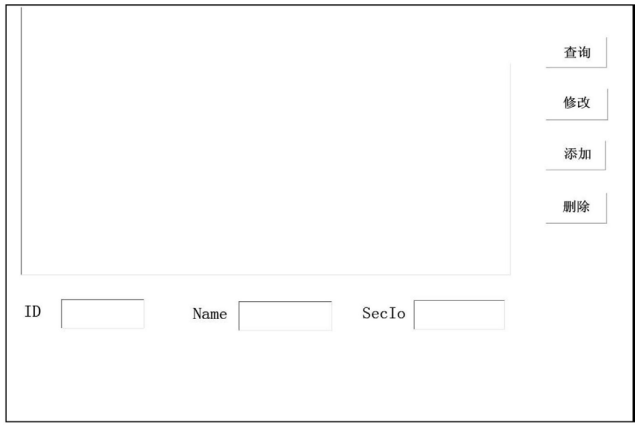

图3 数据库设计界面

在算法执行界面里，人们可以对sale表进行录入操作，并提供最小支持度输入窗口，使得用户可以根据自己的需要对最小支持度进行设置[17]，如图4所示。

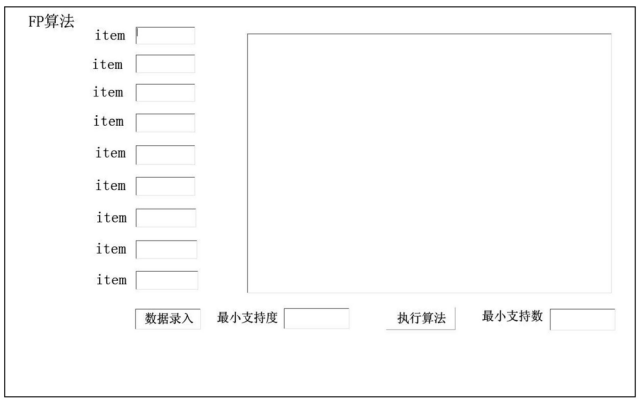

图4 算法执行界面

1.3 系统实现和结果分析

在本系统中，创建了一个简单的登入界面，在界面里显示了用户名和密码，如图5所示。

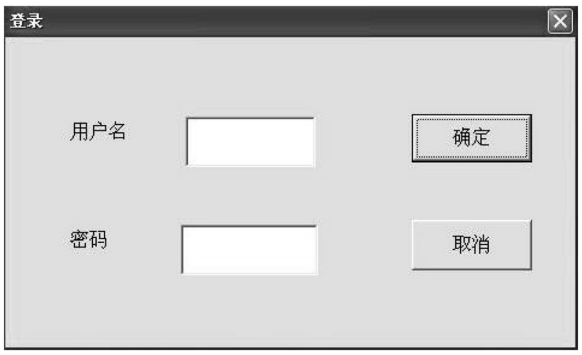

图5 系统实现登入界面

输入的用户名和密码正确后，可以选择对数据库进行操作或者对FP-Tree算法进行操作，这样能方便地对整个程序进行管理，也方便使用，如图6所示。

图6 系统实现选择功能界面

1.3.1 数据库相关界面

进入数据库操作后程序自动弹出如图7所示界面，在该界面上能够实现数据的查询、插入、修改、删除等操作。界面中还包含了现实窗口、序列现实口、生产日期、价格等。这些使得人们更能直观地进行使用。

以查询为例，SecIo是按照商品属性的第几列来查询，如ID对应第一列，Name对应第2列，以此类推。这样可以从数据库中调出人们需要查找的那一项。当人们要查询ID为a的商品时，生成结果如图8所示。

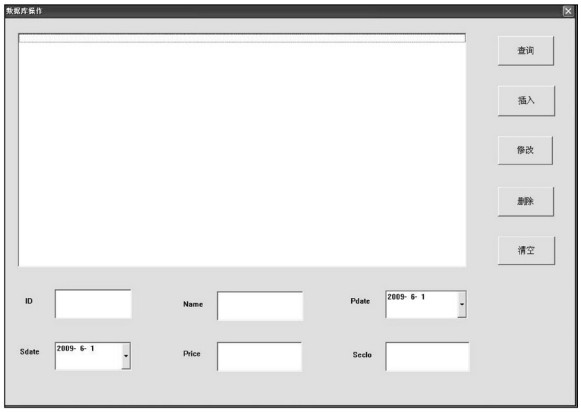

图7 系统实现数据库操作界面

图8 查询功能

在进入FP-Tree算法界面后，能看到如图9所示的界面。其中Item是顾客在一次购买行为中购买的商品（项）的集合。只需在后面空白处输入对应的商品ID，输入完成后点击数据录入按钮即可。重复上述操作完成数据录入。事务查询是对购买商品的整体查询，按第几列查询，并在第几列输入你所要查询的商品ID即可。在执行算法之前只需在“请输入支持度”按钮后面输入所需的支持度即可。支持度定义在0到1之间，事务个数由程序自动生成，它代表了所有顾客总的消费次数，最小支持度为支持度与实务个数的乘积。

1.3.2 FP-GROWTH算法相关界面

在使用中，人们对不同的支持度产生的选项有需求。那么就需要程序可根据我们输入的支持度来输出，这为下一步的决策提供了依据。可根据当时的需要输入相应的支持度，在这里以0.4为例，如图10所示。

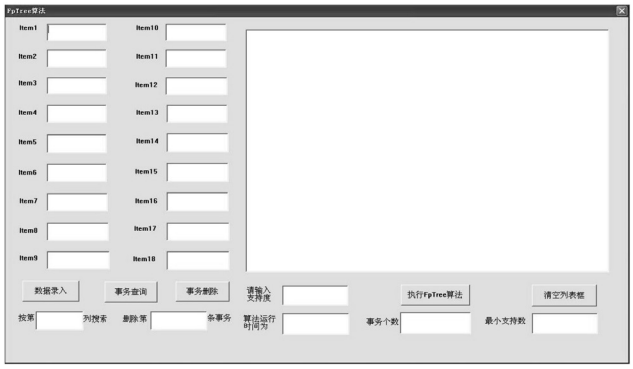

图9 系统实现算法执行

图10 系统实现结果

1.3.3 系统实现结果

数据库中的原始数据如图11所示，在这些数据的基础上人们能方便地得出要预计的结果。

图11 原始数据

从图11能了解到商品和销售的一些基本信息，为了验证程序的正确性，在sale表中给出了3条销售记录。这些记录源于利物浦大学计算机科学系给出的实验原始数据。从图11我们可以直观地发现一个集合db，根据需求输入了一个支持度0.4，在整个数据库中存在着3个实务个数，由于最小支持数是事务个数与支持度的乘积，很容易就得到了最小支持度为1.2，把测得项的支持度与1.2对比，支持数比1.2大的集合为{b，d，db}。根据程序设计要求，把符合要求的最大项输出实现结果如图12所示。

图12 实现结果

2 Apriori算法的性能瓶颈

2.1 对数据库的扫描次数过多

当事务数据库中存放大量事务数据时，在有限的内存容量下，系统I/O负载相当大。对每次k循环，候选集CK中的每个元素都必须通过扫描数据库一次来验证其是否加入LK。假如有一个频繁大项集包含10个项的话，那么就至少需要扫描事务数据库10遍。每次扫描数据库的时间就会非常长，这样导致Apriori算法效率相对低。

2.2 可致使庞大的侯选集的产生

由LK-1产生k-侯选集CK是指数增长的，例如10^4的1-频繁项集就有可能产生接近10^7个元素的2-侯选集。如果要产生一个很长的规则时，产生的中间元素也是巨大的。

2.3 有用数据少

基于支持度和可信度框架理论发现的大量规则中，有一些规则即使满足用户指定的最小支持度和可信度，但仍没有实际意义；如果最小支持度阈值定得越高，有用数据就越少，有意义的规则也就不易被发现，这样会影响决策的制定。

2.4 算法适应范围小

Apriori算法仅仅考虑了布尔型的单维关联规则的挖掘，在实际应用中，可能出现多类型的、多维的、多层的关联规则[18-21]。

两种算法最小支持度设置与执行时间的性能比较如图13所示，当最小支持度分别为1%，0.5%，0.2%，0.1%时，Apriori算法比FP-GROWTH算法运行时间长。当两个算法运行时间相同时，FP-GROWTH算法能够实现比Aprior算法最小支持度更低的运算。从测试结果可以看出，FP-GROWTH算法在挖掘频繁项目集方面的性能要好于Apriori算法。

3 结语

在本课题里，我们对基于FP-GROWTH算法的关联规则挖掘算法进行了理论研究和系统的实现。描述了数据挖掘和关联规则的背景、理论基础以及国内外研究现状。采用了C++，SQL server来实现的系统，在系统中，人们能够对数据库进行基本操作并能执行算法。对于最后结果笔者给出了分析，并与Apriori算法进行了比较。

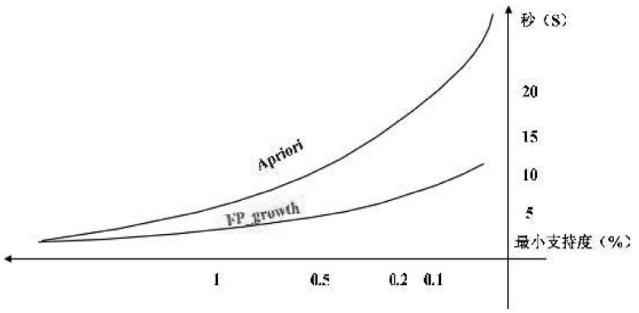

图13 算法对比

由于时间的仓促，系统的界面排版还不够完美。程序主界面较为简单，没有插入图片点缀界面使界面更加好看。数据库设计方面，设定的查询、修改、删除等操作较为繁琐，不够人性化。在当时设计算法界面时，没把顾客购买的项目集和算法执行分开，导致算法执行界面较为复杂。使用者在操作时容易混淆相关操作。界面的布局相当拥挤，这是在当时设计时没有想到的。

虽然系统还有很多需要完善之处，但就系统的设计思路和方法来说还是有其特色的。将会在现有基础上，进一步改进和完善系统，使其可以达到更好的效果，更好地满足用户的使用需求。通过对基于FP-GROWTH算法的关联规则挖掘算法系统的开发，让人们对实际系统的开发有了一个初步的整体概念，真正体会到了软件工程以及面向对象的思想在实践中的指导作用。

[参考文献]

[1]刘乃丽，李玉忱，马磊，等.一种基于FP-tree的最大频繁项目集挖掘算法[J].计算机应用，2005（5）：998-1000.
[2]韩家炜，米歇尔·坎伯. 数据挖掘概念与技术[M].2版.范明，孟小峰，译.北京：机械工业出版社，2001.
[3]朱明.数据挖掘[M].合肥：中国科学技术大学出版社，2002.
[4]HAN J, KAMBER M. Data mining concepts and techniques[J].Data Mining Concepts Models Methods & Algorithms Second Edition，2012（4）：1-18.
[5]李瑞轩，卢正鼎.多数据库系统原理与技术[M].北京：电子工业出版社，2005.
[6]张云涛，龚玲.数据挖掘原理与技术[M].北京：电子工业出版社，2004.
[7]崔立新.约束性相联规则发现方法及算法[J].计算机学报，2000（2）：216-220.
[8]毛国君，段立娟.数据挖掘原理和算法[M].北京：清华大学出版社，2016.
[9]邵峰晶，于忠清.数据挖掘原理与算法[M].北京：中国水利水电出版社，2003.
[10]李冠乾，许亮.CRM数据挖掘中关联规则的应用[J].昆明理工大学学报（自然科学版），2004（1）：113-117.
[11]赵岩，赵慧娟.数据挖掘理论与技术[J].福建电脑，2006（2）：54.
[12]薛慧君.数据挖掘技术及其在电子商务中的应用研究[J].内蒙古农业大学学报（自然科学版），2005（4）：86-90.
[13]杨克俭.数据挖掘及其应用研究回顾[J].福建电脑，2004（7）：9-10.
[14]李菁菁.数据挖掘在中国的现状和发展研究[J].管理工程学报，2004（3）：10-15.
[15]玛格丽特·邓纳姆.数据挖掘教程[M].郭崇慧，田凤占，靳晓明，译.北京：清华大学出版社，2005.

（下转第124页）

3.3 工具研发

随着PMS2.0系统上线运行，系统“应用情况指标”查询模块存在考核细度低于现场要求，功能完善进度慢等问题，影响了实际指标管控、分析工作。为此，安徽公司结合PMS2.0系统指标体系与本地基层实际业务应用情况，研发了“PMS智能管控分析平台”，以精益化管理为主线，以推进PMS2.0系统实用化应用为重点，统计查询各业务指标，提升各地市公司专业人员业务应用系统能力，建立统一、高效、集约的运维检修管理信息化平台，满足执行层、管理层和决策层需要，提升电网运检管理精益化水平。

“PMS智能管控分析平台”针对基础设备台账、生产业务数据、系统运行数据的各类指标进行分析、预警，秉承“以指标促应用，以应用促管理”理念，从月度、季度、年度3个维度对指标进行科学组合，强化指标评价对生产信息化专业管理的支撑作用，合理设置发布周期，科学指导各单位系统实用化推进，加大各单位管理改进和自我提升的关注度。

4 结语

PMS2.0是“三集五大”体系建设中的“大检修”体系内容，支撑了运维检修全过程精益化管理和电网资产的全寿命周期管理，安徽公司通过以上技术上、管理上的一系列措施，有效地提升了系统性能，提高了用户体验。目前，该系统已在安徽公司全面应用推广，系统运行稳定，功能应用可靠，有效支撑了安徽公司现有设备资产的运维检修、全寿命周期管理。

Optimization and application of equipment (asset) operation and maintenance lean management system

Zhang Yongmei, Jia Hui, Tang Yixuan, Yao Zhen, Wang Li
（State Grid AnHui Information &Telecommunication Company, Hefei 230061, China）

Abstract: With the equipment (asset) operation and maintenance lean management system of State Grid Anhui Electric Power Company on the line, achieved a horizontal, vertical multi-system data sharing and business integration, and promoted the level of production management information to a new level. This paper introduces the system tuning in the process of system construction, such as slow data access, graphics caton and poor interface stability, and elaborated the deepen application management methods and the corresponding measures of the PMS2.0 system.

Key words: system structure; performance tuning; “126” management and control; deepen application

（上接第121页）

[16]李敏强，寇纪淞.遗传算法的基本理论与应用[M].北京：科学出版社，2002.
[17]吉根林，帅克，孙志挥.数据挖掘技术及其应用[J].南京师范大学学报（自然科学版），2000（2）：25-27.
[18]唐华松，姚耀文.数据挖掘中决策树算法的探讨[J].计算机应用研究，2000（8）：18-22.
[19]周志华，陈世福.神经网络集成[J].计算机学报，2002（6）：587-590.
[20]李永敏，朱善君，陈湘晖，等.基于粗糙理论的数据挖掘模型[J].清华大学学报（自然科学版），1999（1）：110-113.
[21]糜元根.数据挖掘方法的评述[J].南京化工大学学报，2001（9）：105-109.

Study on association rule mining algorithm based on FP-GROWTH algorithm

Chen Yin
（Nanjing Howso Technology Co., Ltd., Nanjing 210012, China）

Abstract: The data of Internet world is doubling every year, but the useful information for the user seems to decrease, users are drowning in the ocean of data, although search engines such as Google can help users find the information they need, but the correct rate and the recall rate is not satisfactory. Data mining is a new technology used in decision support since 1990s. The FP-GROWTH algorithm scan only twice databases. It does not use the candidate set, but directly compresses database into a frequent pattern tree, finally the association rules generated by this tree. While this research studies FP-GROWT H algorithm theory, it implements the system of a simple algorithm demonstrating. The system includes the implementation of the algorithm, the database modification, query, delete operations. Finally, the FP-GROWTH algorithm and the Apriori algorithm are compared in this paper.

Key words: data mining; association rule; FP-GROWTH algorithm; candidate set; frequent pattern tree