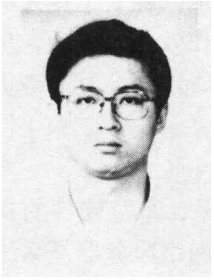

王建宇

数据挖掘在铁路机务段MIS中的应用

王建宇 王荣平 朱卫东

摘 要 以铁路机务段现有的数据库为基础，利用数据挖掘的方法，研究提出了关联规则挖掘算法，并将其应用于铁路机务段管理信息系统，通过一个原型的实践表明，挖掘的关联规则和知识对提高企业MIS的管理水平有显著作用，说明了进一步研究的方向。

关键词 数据挖掘(Data Mining)，关联规则(Association Rules)，数据库

Application of Data Mining to MIS of Railway Locomotive Depot

Wang Jianyu Wang Rongping Zhu Weidong

(Information Center of Northern Jiaotong University, Beijing, 100044)

Abstract: The writer discussed the mining of association rulés in detail, based on theory of data mining and present databases . The study supplied a prototype to practical applications of data mining in MIS. In addition, the paper mentioned the farther direction of study.

Keywords: Data Mining, association Rules, Database

1 引言

大型关系数据库中存在着大量的数据，以往我们通过OLTP(联机事务处理)进行查询、汇总来对存在于数据库中的数据进行简单的处理，处理过程和结果比较简单，不能满足企业的需求。进而又出现了OLAP(联机分析处理)，它专门设计用于支持复杂的分析操作，侧重对决策人员和高层管理人员的决策支持，可以按分析人员要求快速、灵活地进行大数据量的复杂查询处理，并且以一种直观易懂的形式将查询结果提供给决策人员，准确掌握企业(公司)的经营状况，了解市场需求，制定正确方案，增加效益[3]。但是，以上两种处理只是对数据库中数据静态地反映，实际上在数据库中隐藏着许多鲜为人知的信息，原有的处理工具已不能满足企业领导者的需要。那么，怎样从中提取出高质量的信息(预测性)是摆在数据库研究人员面前的一个新的课题，而这正是数据挖掘研究的内容。

数据挖掘(Data Mining)是数据库技术、人工智能技术、统计学相结合的产物，也是目前国际上数据库领域和决策支持研究领域的最前沿方向之一。在信息技术高度发达的今天，数据挖掘帮助人们从浩如烟海的数据海洋中将最有价值的信息挖掘出来，在这方面数据挖掘显示了巨大的潜力，目前，这一研究领域正在蓬勃的发展，美国国家科学基金会(NSF)已将其列入90年代最有价值的数据库研究项目。数据挖掘的主要目标有关联规则的发现、序列规则的发现、分类规则的发现等。采用的技术有规则归纳、决策树、粗集理论、神经网络、遗传算法等。数据挖掘的主要应用领域有金融、医疗保健、市场业、零售业、制造业、司法、工程与科研等部门。我们开发的机务段管理信息系统在运行过程中，积累了大量的关于机车维护和检修的数据，这些数据中同样也隐藏着一些知识和规则，那么怎样挖掘出

王建宇 北方交通大学信息中心 在读硕士研究生 100044 北京市
王荣平 北方交通大学计算中心 在职硕士研究生 100044 北京市
朱卫东 北方交通大学计算中心 副教授 100044 北京市

这些潜在的知识来提高机车检修质量，减少机车运行中的隐患，成为我们的研究重点。

2 关联规则

2.1 关联规则的定义

关联规则的发现自 R. Agrawal 等人于 1993 年提出来后[1]，一直是数据挖掘研究中一个重要目标。关联规则就是数据库中隐藏在数据间的相互关系。例如：一个市场销售情况数据库，从中发现有 80%买了麦片粥和糖的顾客，同时也买了牛奶；再例如，一个人口普查的数据库，我们可以发现 60%去年工作的人中的工资收入小于平均水平。

对关联规则的定义如下：

定义 1 $I=\{i_1,i_2,\cdots,i_m\}$称为数据项集，其中 i 为数据项.

定义 2 $D=\{t_1,t_2,\cdots,t_n\}$称为事务集，其中每一个事务 t_i 对应一个数据项集，$t_i\subseteq I$，
每个 t_i 都有一个唯一的标识符 TID 与其对应.

定义 3 $I=\{i_1,i_2,\cdots i_k\}$称为 K－数据项集，K 是 I 中数据项的个数.

定义 4 数据项 I 的支持度 support(I)是事务集 D 中包括数据项 I 的事务，在 D 中的百分比，用公式表示为 $\text{support}(I)=\|\{t\in D\mid I\subseteq t\}\|/\|\{t\in D\}\|$

定义 5 关联规则是形如 $I_1\Rightarrow I_2$ 这样的产生式，其中 $I_1,I_2\subset I$，且 $I_1\cap I_2=\Phi$。

定义 6 关联规则 $r:I_1\Rightarrow I_2$ 的支持度：$\text{support}(r)=\text{support}(I_1\cup I_2)$.

关联规则 $r:I_1\Rightarrow I_2$ 的置信度：$\text{confidence}(r)=\text{support}(I_1\cup I_2)/\text{support}(I_1)$

2.2 关联规则挖掘描述

一般地，在关系数据库中挖掘关联规则的过程是这样的，首先，给出最小支持度(minisupport)、最小置信度(minconfidence)，然后找出所有大于等于给出的最小支持度和最小置信度的关联规则，这个问题可以分为下面两个子问题：

(1)在数据库中找出所有大于最小支持度的频繁数据项集，频繁数据项集也被称为大数据项集。

(2)对每一个找到的频繁数据项集 I_1，生成关联规则 $I_2\Rightarrow I_1-I_2\mid I_2\subset I_1$，当然此规则的置信度 confidence 应当大于等于最小置信度。

其中第一步是，对于一个含有项目集个数为 m 的事务数据库来说，频繁项目集的可能解空间为 2^m-1。第一步的目的就是从这些可能解空间中将频繁项目集找出来，算法的复杂性充分体现在第一步上。下面一节我们给出具体算法及相关的性质：

2.3 关联规则挖掘算法[2]

(1)频繁项目集的计算

首先，我们引入若干记号。

由 k－项目集构成的集合称为 k－项目序列集，L_k 记由频繁 k－项目集构成的集合，C_k 记由候选 k－项目集构成的集合。

其次，我们引入定理如下：

定理 1 任意频繁 k－项目集的任意子集均是频繁集。

证明：首先证明 k－项集 r 的任意子集均为频繁项目集。根据条件以及 r 的构造方法，去掉 r 中任何 1 个项所得的($k-1$)项集 $s\in L_{k-1}$，即 s 是大($k-1$)项集，其任意子集亦为频繁项目集，故 r 中去掉任意多个项得到的非空子集为频繁项目集。其次，证明产生全部满足 C_k 条件的 k－项集。令 r 是任意一个满足 C_k 条件的 k－项集，分别去掉 r 的第 k－项和第 $k-1$ 项得到两个 $k-1$ 项集 $u=\{r[1],r[2],\cdots,r[k-2],r[k-1]\}$，$v=\{r[1],r[2],\cdots,r[k-2],r[k]\}$，$u\in L_{k-1}$，$v\in L_{k-1}$，由 L_{k-1}中的 u 和 v 可构造 r，即产生 C_k。

频繁项目集的发现是一种渐进的方法，即按照项目集的长度，从发现频繁 1－项目序列集开始，逐次增加项目集的长度。

算法 1：(Apriori frequent itemset discovery)

```
① L1←{Frequent 1−itemsets};
② for (k←2; Lk−1;k++) do begin
③   Ck←Apriori-Gen(Lk−1);        //生成候选 K−项目集 Ck
④   forall transactions t∈D do begin
⑤   Ct←Subset(Ck,t);             //判断 Ck 是否属于事务 t
⑥   forall candidates c∈Ct do
⑦       c.count++;
⑧   end
⑨   Lk←{c∈Ck| c.count≥minsupport};
⑩ end
⑪ answer←kLk;                    //生成频繁项目集
```

Apriori-Gen Function

```
① insert into Ct
② select p[1], p[2],…, p[i−1],q[i−1]
③ from Ln−1p,Ln−1q
④ where p[1]=q[1],…,p[i−2]=q[i−2], p[i−1]<q[i−1];
⑤ forall candidate itemsets c∈Ct do begin
⑥     forall(i−1)-subsets s of c do begin
⑦         if (s∉Ln−1) then
⑧             delete c from Ct;
⑨     end
```

⑥～⑨对 C_k 中的任一候选 c，如果 c 中存在一个

不属于 C_{k-1} 的长度为 $k-1$ 的子序列，那么就从 C_k 中删除该候选 c。

```
⑩ end
⑪ Answer←U{c∈C_k};
```

(2) 规则生成

规则生成是关联规则挖掘中相对比较容易的一步。对每个频繁项目集 L_k 输出形如 $(l-a)\rightarrow a$ 的规则，其中 a 是 l 的非空子集且满足 support(l)/support($l-a$)minconf。

令 a' 为 a 的任意非空子集，$a'\subset a$，有 support($l-a'$)support$\leqslant$($l-a$)，规则 $(l-a')\Rightarrow a'$ 的置信度不小于 $(l-a)\Rightarrow a$ 的置信度. 因而，若 a' 不能作为规则的结论，则 a 亦不能。反之，若有 $(l-a)\Rightarrow a$，则必有 $(l-a')\Rightarrow a'$。

例如，若规则 $AB\Rightarrow CD$ 成立，则 $AB\Rightarrow CD$，$ABD\Rightarrow C$ 亦成立. 据此，对于任意频繁项目集 L，我们首先生成所有仅含 1 个项为其结论的规则，然后根据这些规则的结论利用 Candi_gen 生成包含 2 个项的频繁项目集，作为 L 的可能结论，并计算其置信度，得到所有结论包含 2 个项的规则，如此下去。生成规则的完整算法如下：

```
① forall frequent k-item sets l_k∈L_k, k_2≥2 do begin
②   H_1←{l_k 的长度为 1 的子集};
③   forall h_1∈H_1 do begin
④         confidence←support(l_k)/support(l_k-h_1);
          //规则 r:(l_k-h_1)⇒h_1 的置信度
⑤         if (confidence≥minconfidence) then
⑥               AR←AR∪{r:(l_k-h1_1⇒h1 ;
⑦         else H_1←H_1-{h_1 ;
⑧   end
⑨   call gen-rules(l_k,H_1);      //H_1=从 l_k 中产生的 1-项目集
⑩ end
⑪ procedure gen-rules(l_k:frequent k-itemset, H_m:set of m-item consequences);
⑫ if (k>m+1) then do begin
    forall h_{m-1}∈H_{m-1} do begin
      confidence←support(l_k)/support(l_k-h_{m-1});
      if (confidence≥minconfidence) then
        AR←AR∪{r:(l_k-h_{m-1})⇒h_{m-1}};
      else delete h_{m-1} from H_{m-1};
        end
        call genrules(l_k,H_{m-1})
      end
    end;
```

3 在机务段管理信息系统中的应用

机车的检修质量是机务段非常重视的问题。目前我们国家的铁路线还是非常繁忙的，在运输过程中机车是不能发生故障的，否则会使这条线的运输陷入停滞，会给国家造成很大的损失。所以，尽量避免故障车是整个铁路运输中的一个非常重要的问题。但是在目前的情况下，有很多情况是检修工作无法发现的，譬如，如果某些配件本身的质量有问题，那么虽然经过换新处理，但实际上这些零件很快就会出问题的，但这时机车还没有到下一次检修时间。如果这台机车现在正在担任运输任务，很有可能会在半路抛锚，带来不应有的损失。所以，能否利用我们已有的检修记录，配件的领取记录来找出它们之间的关系，预测是否有配件存在质量隐患，就成为我们的一个研究重点。如果能找到这种关系，这对机务段提高机车安全运行将起到很大的作用。

在机务段管理系统中，存储了大量与修车有关的信息，如检修班组领取配件记录，配件定期维修的记录，而且这些信息的数量比较大，若用人工的方法去发现这些数据中隐藏的规则是不可能的，但是这些数据正是数据挖掘的一个很好的基础。我们可以通过适当的转换将其生成一个事务数据库。在这个基础上，我们将采用上面提出的方法进行挖掘。

我们编制的原型采用的编程环境为客户端采用 PB，服务器端为 UNIX，ORACLE。因为不是每一个数据库都可以进行挖掘，所以系统首先必须对挖掘的数据进行转换，转换需将离散的或连续的属性值转换为挖掘算法可以处理的布尔型属性值，然后转换为可以进行挖掘的事务数据库，这个事务数据库的每一条记录包含有唯一的 TID，其余的内容为转换过后生成的项目集中的项目。然后采用上面我们介绍的算法进行挖掘。实验结果表明挖掘的结果还是比较好的，能够反映出隐含的一些规律。

4 结束语

管理信息系统做为企业生产和管理的工具，被日益应用到各行各业，但其只能做为对已有数据的静态反映，不能反映出大量数据里隐藏的知识和规则，也就是说不能提供给决策者更进一步的管理策略，但数据挖掘的出现在一定程度上弥补这方面的空白，它能反映出大量数据中的潜在的知识和规则，在这个基础上，企业的管理水平可以上升到智能的管理水平，这也充分体现了数据挖掘是数据库技术和人工智能相结合的产物。

鉴于以上原因，我们将数据挖掘技术应用到铁路机务段管理信息系统中，实践证明挖掘出的知识和规则对提高企业的管理水平有显著作用。当然，在关联规

李英女

铁路客运信息查询算法

李英女 郑国雄

摘 要 铁路客运信息查询算法是建立在数据库基础上的。为用户提供全国范围内任意两个车站之间合理的乘车方案，该算法需要建立的数据库包括：车次表、车站表、时刻表和相关局表，通过这些基本数据库，使用中转算法可以产生中转站表，从而构成铁路客运信息查询算法的基础。当用户任意给定两个车站时，查询算法能在短时间内给出合理的乘车路线、乘车车次、乘车时间以及中转位置等信息，用户可以利用这些信息指导自己的铁路旅行。

关键词 客运 数据库 查询 算法

Inquiry Arithmetic for Railway Passenger Transport Information

Li Yingnu Zheng Guoxiong

(Beijing Institute of Tracking and telecommunication Technology, 100094)

Abstract: The inquiry arithmetic for railway passenger transport information is based on the database, and its function is to serve the user an available scheme for them to transfer between each two railway stations all over the country. The databases for this method include : table of train numbers, table of railway stations, table of schedule and table of correlative railway bureau. Using these original tables, we can get transfer tables with Transfer Arithmetic, which are the foundation of the inquiry arithmetic for railway passenger transport information. When given any two railway stations , the inquiry arithmetic can offer a reasonable route, train numbers ,starting times and transfer stations ect. . With these information, user can guide one's railway travel life effectively.

Keywords: railway passenger transport, database, inquiry, arithmetic

1 引言

铁路运输具有运能大、速度快、能耗小、安全可靠、对环境污染小、运费低，选择铁路旅行是目前我国绝大多数人和许多国外来华旅游团主要的中长途旅行方式。

我国正在运营的铁路客运系统中，有8种列车，分别是：准高速列车、快速列车、旅游列车、特别快车、普

李英女 北京跟踪与通信技术研究所 工程师 100094 北京市
郑国雄 北京航天指挥控制中心 在职博士研究生 100094 北京市

则挖掘方面我们还有许多问题需要进一步研究，如怎样提高算法的效率，怎样对挖掘出的规则进行进一步的剪裁和总结等等。

5 参考文献

1 R. Agrawal, T. Imielinski, and A. Swami, Mining association rules between sets of items in large databases.. In Proc. 1993 ACM—— SIGMODInt. Conf. Management of Data, pp. 207～216, Washington, D. C. , May 1993.

2 N. Pasquiter, Y. Bastide, R. Taouil and L Lakhal. Efficient Mining Of Association Rules Using Closed Itemset Lattices. Information Systems Vol. 24, No. 1, pp. 25～46, 1999

3 王珊. 数据仓库技术与联机分析处理. 北京：科学出版社，1998

（责任编辑：赵存义 收稿日期：2000－01－12）