第 33 卷 第 11 期 计 算 机 工 程 2007 年 6 月
Vol.33 No.11 Computer Engineering June 2007
·软件技术与数据库· 文章编号：1000—3428(2007)11—0034—03 文献标识码：A 中图分类号：TP311

# 海难事故的数据挖掘

于卫红[1]，贾传荧[2]
(1. 大连海事大学经济与管理学院，大连 116026；2. 大连海事大学航海学院，大连 116026)

**摘 要：**分析了建立海难数据仓库的意义，提出了海难数据仓库的雪花模型，对 Aprioir 算法进行了改进，用改进后的算法实现了海难数据的关联规则和频繁模式挖掘，用改进的有向图方法实现了关联规则的可视化表示。结果表明，利用数据挖掘技术对海难历史数据作深层次分析，克服了传统统计分析方法的局限性，可挖掘出大量的知识，为以后的航海安全提供借鉴。
**关键词：**Apriori；关联规则；数据挖掘；海难；可视化

# Data Mining in Shipwreck Data Warehouse

YU Weihong[1], JIA Chuanying[2]
(1. Economy and Management College, Dalian Maritime University, Dalian 116026;
2. Navigation College, Dalian Maritime University, Dalian 116026)

**【Abstract】**This paper analyzes the meaning of data mining in shipwreck data, improves the Apriori algorithm and applies it into finding frequent patterns of shipwreck data warehouse which is organized as a snowflake schema while improving direct rule graph to visualize the association rules. Research result shows that data mining technology for further research on historical shipwreck data can overcome the limitations of the traditional statistics and analysis and can mine a lot of knowledge so that some references can be provided for future navigation safety.
**【Key words】**Apriori; Association rule; Data mining; Shipwreck; Visualization

目前，美国、澳大利亚、加拿大等国都建立了自己的海难数据库。建立海难数据库是一个庞大的工程，我国海事统计数据存在着不规范、不完整的现象，相关的事故数据缺少必要的现场动态安全信息。海事信息化建设需要建立各种数据库，如全国统一的船舶数据库、海难数据库。海难数据库的主要作用是记录每次海上事故发生、处理情况；实现分地区、分时间、分种类对海上事故进行查询、分析、统计；实现应急力量的统一管理；实现应急方案的智能化处理。

海难事故的历史数据是今天研究航海安全的重要财产，但是只有真正揭示出隐藏在原始数据背后的信息，这些数据才有意义。目前我国对海难事故的分析主要限于查询、报表、联机应用分析等传统的分析手段。海难事故数据库中的数据是多维、稀疏的，这是因为与之相关的影响因素很多，如人的因素、船自身结构、天气、航道信息等。仅用数理统计方法处理高度复杂而又信息不全的系统，处理效率低下，提供的信息不适用。只有通过多模型结合的学习，才能挖掘出有用的事故信息，发现内在的客观规律。

此外，目前海事部门的安全工作仍是经验性的，对象比较单一，没有将系统分析方法引入安全工作，仍处于被动的状态，因而事故的预测、预防能力差。

数据挖掘是人工智能中的一种有效的新技术，它包括多种理论模型，可以克服传统方法解决具体问题时存在的局限性，同时挖掘的知识可以发展、丰富相关的专家决策信息系统，逐步形成相应的安全体系，为事故的分析和预测提供更广阔的平台。

## 1 数据源与数据建模

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。

我国尚缺乏一个完善的海难数据库，本文比较了其他国家的海难数据库，设计了海难数据仓库的雪花模型。如图 1 所示。目前最流行的数据仓库数据模型是多维数据模型，主要包括星型模型和雪花模型。星型数据仓库包括一个大的包含大批数据和不含冗余的事实表和一组小的维表，每维一个。雪花模型是星型模型的变种，不同的是雪花模型的维表是规范化的、减少了冗余、易于维护和存储、增加了应用程序的灵活性、易于实现动态 SQL 生成。

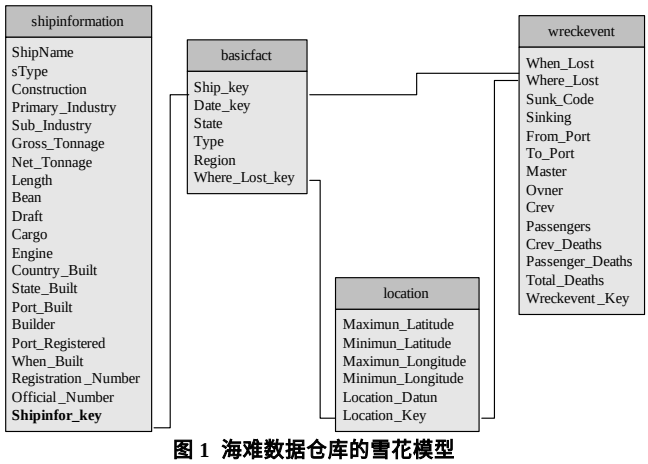


**图 1 海难数据仓库的雪花模型**

**作者简介：**于卫红(1972－)，女，副教授，主研方向：多Agent系统，数据挖掘，电子商务，智能决策支持系统；贾传荧，教授、博导
**收稿日期：**2006-06-30 **E-mail：**yuwhlx@163.com

图 1 所示的海难数据仓库雪花模型中包含两种表：事实表(表 basicfact)和维表(表 shipinformation、表 location 和表 wreckevent)。事实表描述海难事故的基本事实，表 shipinformation 描述难船的基本信息，表 location 描述海难地点的位置信息，表 wreckevent 描述事故具体信息，如人员伤亡、天气状况等。各表都是满足 3NF 的规范化表格。

## 2 数据提取、转换、清理

作为方法研究，我们从澳大利亚国家海难数据库中查询出发生在 1900 年-2005 年之间的海难数据，大约有 5 000 多艘难船的信息。由于无法直接存取原有的数据库，所有的查询结果都是以 HTML 形式存储的网页文件，不具有规范化和格式化，无法直接将数据存储到数据库中，因此使用批处理的方法将 5 000 多个类似的网页文件下载到本地， 开发一个批处理转换工具将所有网页文件中的 HTML 标记去掉，提取有用的数据用 insert into 命令存入到图 1 所示的数据库中。

## 3 数据挖掘的实现

### 3.1 Apriori 算法及其改进

关联规则挖掘发现大量数据中项集之间有趣的关联或相关联系。可对关联规则描述如下：设$I=\{i_1,i_2,..,i_m\}$是项集，任务相关的数据D是事务集，其中每个事务T是项集，使得$T\subseteq I$。设A是一个项集，且$A\subseteq T$。关联规则是如下形式的逻辑蕴涵：$A\Rightarrow B$，$A\subset I$，$B\subset I$，且$A\cap B=\Phi$。关联规则具有如下 2 个重要的属性：

(1)支持度：P(A∪B)，即 A 和 B 这两个项集在事务集 D 中同时出现的概率。

(2)置信度：P(B｜A)，即在出现项集 A 的事务集 D 中，项集 B 也同时出现的概率。

同时满足最小支持度阈值和最小置信度阈值的规则称为强规则。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，即产生强规则的问题。所有支持度大于最小支持度的项集称为频繁项集，简称频集。

关联规则挖掘主要通过 Apriori 算法实现，流程如图 2 所示。该算法主要解决 2 个问题：

(1)产生候选频繁集；

(2)计算候选项的支持度。

主要完成 2 个动作连接和剪枝。算法中D表示事务数据库，min_sup为最小支持度，L是D中的频繁项集，$L_k$是频繁K项集。为找$L_k$，通过$L_{k-1}$的自连接产生候选K项集合，该集合记作$C_k$。$C_k$是$L_k$的超集，即它的成员是频繁的，但所有的频繁K项集都包含在$C_k$中。

自连接可用如下 SQL 语句来实现：

Step 1　self-join $L_{k-1}$

INSERT INTO $C_k$

SELECT p.item1, p.item2, …, p.itemk-1, q.itemk-1

FROM $L_{k-1}$ p, $L_{k-1}$ q

WHERE p.item1=q.item1, …, p.itemk-2=q.itemk-2, p.itemk-1 < q.itemk-1

剪枝的算法描述如下：

Step 2　pruning

For each itemset c in $C_k$ do

For each (k-1)-subsets s of c do if (s is not in $L_{k-1}$) then delete c from $C_k$

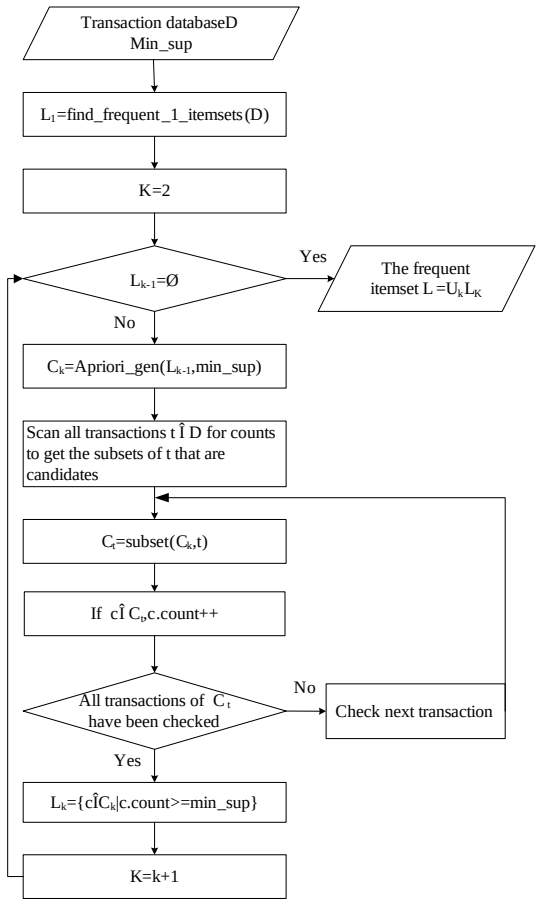


**图 2 Apriori算法流程**

Apriori 算法的缺陷主要表现在：

(1)事务数据库扫描次数过多；(2)候选项集数量过大；(3)侯选项支持度的计算工作量过大。

相应地，通常有 3 种改进策略：

(1)减少事务数据库的扫描次数；(2)压缩候选项集；(3)简化支持度的计算。

基于临时表的 Apriori 改进算法致力于减少扫描次数，主要依据以下 2 个事实：

(1)对于已知规模的事务数据库 D，任意一个项集 I 的出现频繁度与规模小于｜I｜的事务无关。所以在第｜I｜次扫描数据库 D 时，可以删除规模小于｜I｜的事务记录。

(2)K 候选项集中不包含任何(k-1)项集的项集一定不是频繁项集，因此 K 次扫描时可以将这样的事务记录立即删除，从而减少了下次需要扫描的记录数。

用临时表完成频繁项集的选择。首先把(k-1)项集中的第 1 个项集添加进临时表中；然后把最后一项不同的其他项集添加进临时表，生成 K 项集，计算其频度，若频度大于最小频繁度，则生成该频繁项并保存，否则删除。依此循环，直至生成所有的频繁项。

删除临时表中的不满足条件的项集的函数如下：

procedure impr_del($D_{k-1}$,$L_{k-1}$)

for all items $l_1\in D_{k-1}$

if ｜$l_1$｜<K or $L_{k-1}\not\subset l_1$ then delete $l_1$

return Dk

### 3.2 海难数据库频繁模式与关联规则挖掘

应用上述改进的 Apriori 算法，对以雪花模式组织的多维海难数据进行频繁模式和关联规则挖掘。如挖掘船舶类型、

建造材料与事故类别的频繁模式；海难地点与事故类别的频繁模式等。这些属性分别属于不同的维表，从 shipinformation 维表中抽取 sType 和 construction 属性，从 wreckevent 维表中抽取 sunk_code 属性，使用连接将它们组织成新的一维表。通过分组计算出在海难数据库中船舶种类有 40 多种，如 yacht、fishingboat 等，除不明原因外，海难种类有 7 种，如 burnt、foundered 等。

假设最小支持度 min_sup 为 2%，信任度大于 60%，应用改进的 Apriori 算法挖掘出满足该条件的频繁模式，如表 1 所示。表中 supa 表示(stype, construction)⇒sunk_code 的支持度，supb 是(stype, construction)的支持度。Confidence= supa/supb。

**表 1 属性 Stype，Construction，Sunk_Code 间的频繁模式**

| stype | construction | Sunk_code | supa | supb | confidence |
|---|---|---|---|---|---|
| Launch | Wood | Wrecked | 4.08% | 4.30% | 94.74% |
| Lugger | Wood | Wrecked | 4.98% | 7.25% | 68.75% |
| Schooner | Wood | Wrecked | 3.62% | 5.78% | 62.75% |
| Steamer screw | Iron | Wrecked | 3.17% | 3.51% | 90.32% |
| Steamer screw | Steel | Wrecked | 8.95% | 9.85% | 90.80% |
| Steamer screw | Wood | Wrecked | 7.70% | 8.38% | 91.89% |
| Steamer screw | Wood carvel | Wrecked | 4.64% | 4.87% | 95.35% |

从表 1 中可以得出 7 个关联规则，如：

(launch, wood) ⇒wrecked, confidence=94.74%

(lugger, wood) ⇒wrecked, confidence=68.75%等。

## 4 关联规则挖掘结果的可视化表示

可视化是采用图形、图表等易于理解的方式表达数据挖掘结果。在关联规则的可视化表示中，至少要考虑 5 个要素：(1)规则前件；(2)规则后件；(3)前件与后件之间的关联；(4)规则的支持度；(5)规则的可信度。

目前，2 种最基本也是最常用的关联规则可视化方法是二维矩阵表示法和有向图表示法，它们各有优缺点。

二维矩阵表示法被证明是对一对一的关联规则进行可视化的最有效的技术之一，已经在不同的学科领域得到长期广泛的应用。但该方法可视化多对一的关联规则时，呈现出一定的局限性。如在二维矩阵中，很难区分出关联规则( A+B →C )和(A→C and B→C)的不同。

有向图表示法克服了二维矩阵法的缺点。有向图中的每一个结点都代表一个独立的项。连接 2 个结点的边表示关联。在本文的可视化结果中，除了表示出规则前件、后件与关联外，还突出了支持度与可信度的大小。具体做法为：

(1)用实心圆表示规则后件，实心矩形表示规则前件，带箭头的实线表示关联；

(2)将所有支持度从大到小排序，矩形的大小表示支持度的强弱；

(3)将所有信任度从大到小排序，带箭头实线的宽度表示信任度的大小。

表 1 所示的关联规则的可视化结果如图 3 所示。

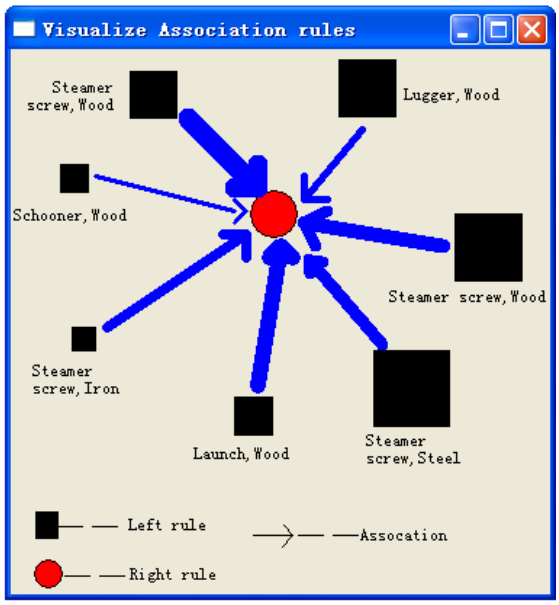


**图 3 表 1 中关联规则的可视化表示**

## 5 结论

海难事故数据是多维的，并在不断增加。用传统的统计分析方法分析它们必然存在局限性。在我国，数据仓库与数据挖掘已被广泛应用到商业领域，但这一新技术却很少被用于航海领域，因此应该尽快建立起健全完善的海难数据库，并使用数据挖掘等先进技术对海难历史数据作深层次研究，从而了解事故真相，挖掘大量知识，为以后的航海安全提供借鉴。本文构建了海难数据仓库的雪花模型，在改进 Apriori 算法的基础上，对海难数据作了关联规则挖掘，并提出了关联规则挖掘结果可视化表示的新方法。

### 参考文献


1 Polese G , Troiano M, Tortora G. A Data Mining Based System Supporting Tactical Decisions[C]//Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering. 2002.

2 Borgelt C, Kruse R. Induction of Association Rules: Apriori Implementation[C]//Proc. of the 14th Conf. on Computational Statistics, Berlin, Germany. 2002.

3 Lee Chang-Hung, Chen Ming-Syan, Member S, et al. Progressive Partition Miner: An Efficient Algorithm for Mining General Temporal Association Rules[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4).

4 吉根林，韦素云，曲维光. 基于平行坐标的关联规则可视化新技术[J]. 计算机工程, 2005, 31(24): 87-89.


（上接第 22 页）

### 参考文献


1 王永吉，陈秋萍. 单调速率及其扩展算法的可调度性判定[J]. 软件学报, 2004, 15(6): 799-814.

2 Lamie W. Preemption-threshold[Z]. http://www.rtos.com/page/imgpage.php?id=210.

3 Wang Yun, Saksena M. Scheduling Fixed-priority Tasks with Preemption Threshold[C]//Proceedings of the 6th Intl. Workshop on Real-time Computing Systems and Applications, Hong Kong. 1999.

4 Liu J W S. Real-time System[M]. Beijing: Higher Education Press, 2000.

5 George L, Rivierre N, Spuri M. Preemptive and Non-preemptive Real-time Uni-processor Scheduling[R]. Inria, Rocquencourt, France: Technical Report 2966: 1996-09.