一种企业能耗预警相关性分析的时空挖掘算法①

郝 平[1,2]，杨剑峰[1]，尹 华[3]

[1](浙江工业大学 计算机科学与技术学院，杭州 310032)
[2](浙江绍兴东越科技有限公司，绍兴 312000)
[3](浙江绍兴市科技信息研究院，绍兴 312000)

摘 要：为了挖掘工业企业能源监测点之间的预警关联规则，寻求生产过程中可能存在的能耗相关性，针对工业企业综合能耗具有的时间特征和空间特征，本文提出了一种新的基于结构化项集的时空挖掘算法，以解决传统的挖掘算法遇到的诸多瓶颈。该算法对经典的 Apriori 算进法行了改进，引入时间维度和空间维度，可在各个时间段和不同的空间层中，挖掘出更多的能耗关联知识。该算法通过引入时间维度和空间维度，建立最小能耗监测单元，构架分时分层的数据挖掘策略，避免传统的挖掘产生过多的候选项集。这种算法应用于国内一家大型企业工业生产过程的能耗数据分析，取得了较好的实际效果。

关键词：能耗；数据挖掘；Apriori 算法；结构项集

Time and Space Mining Algorithm for Correlation Analysis of Energy Consumption Warning

HAO Ping[1,2], YANG Jian-Feng[1], YIN Hua[3]

[1](School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310032, China)
[2](DongYue S&T Co. Ltd, Shaoxing 312000, China)
[3](Science and Technology Information Institute, Shaoxing 312000, China)

Abstract: This paper proposes a structured item-sets time and space algorithm, which is used to find warning association rules between energy consumption monitoring points, according to the time characteristic and spatial characteristics of overall energy consumption in industrial enterprises, and the new algorithm can solve many difficulties that tradition mining algorithm encountered. The new algorithm has been improved on the basis of the classical Apriori algorithm. Introducing the time dimension and space dimension, a new algorithm can dig more associated knowledge during different time periods and in different spatial layers. Through establishing minimum energy monitoring unit and adopting time-sharing and hierarchical data mining strategy, the new algorithm avoids producing excessive candidate sets. The improved algorithm has been applied in energy data analysis of production process of a large industrial enterprise in the domestic and achieved satisfying practical results.

Key words: energy consumption; data mining; Apriori algorithm; structured Utem-sets

随着国际能源的紧缺和能源价格的上涨，能源管理成为近年来的一个研究热点。在最近的几年内，大多数研究侧重通过传统的单维单层挖掘算法对工业能源数据进行挖掘，而通过时间维度和空间维度对流程企业生产过程能耗相关性分析和数据挖掘研究不多。

在已查阅的文献中，文献[1]、[3]、[4]提出一些改进的挖掘算法。这些改进的算法在单维的工业数据挖掘中取得了较好的实验效果。随着工业用能的多样化和工业生产组织管理结构化，工业企业的能耗挖掘也具有多维度多层次的特征。文献[5]~[7]提出一些的多维多层挖掘算法。这些挖掘算法在实验仿真中取得较好的效果。在实际应用中，工业企业的用能又会体现出时间的特征，即不同生产时间段内，工业企业的用能情况会出现较为明显的差异。

① 基金项目:国家科技部项目(2009GJC200001)
收稿时间:2011-10-08;收到修改稿时间:2011-12-07

为此，在错综复杂的工业生产系统中，企业能耗研究者迫切希望通过一种实用的挖掘算法对分布广泛的能耗异常点进行关联挖掘，以寻求生产系统内部尽可能多的能耗相关点，并希望挖掘出的关联知识能够灵活地适用于不同的生产时间段。由文献[5~7]多层挖掘思想的启发，本文总结工业企业综合能耗具有的时间特征和空间特征，提出了一种基于结构化项集的时空挖掘算法，该算法引入了时间维度和空间维度，对经典的 Apriori 挖掘算法作了改进，并以某一大型印染企业作为试点，取得了较好的使用效果。

1　问题描述与分析

1.1 问题描述

在现实工业环境中，重要的监测数据往往随着时间、环境、温度等等的外部因素的改变而呈现不同的分布趋势，考虑外部因素的敏感特性，要在各个监测点中挖掘它们之间的关联性，采用传统挖掘算法将遇到诸多瓶颈。例如，当项集 I 中的各项关联关系具有时间特征时，通过设置最小支持度阈值 $(MinS)$ 和最小置信度阈值 $(MinC)$。采用经典的 Apriori 算法对事务集 D 进行挖掘，很可能产生以下两方面的问题：

(1) 需要重复地扫描事务集：当对超大型事务集 D 的进扫描时，算法的运行效率将大打折扣[8]。

(2) 遗漏重要的关联知识：在某一时期内敏感的关联规则，很可能由于小于事务集 D 最小支持度阈值而被遗漏。

另一方面，随着生产自动化和工业环境的复杂化，为准确掌握监测点的异常情况，工业企业中一个重要监测点往往配备了多类监测仪表。这样的监测点自身内部同样有着复杂的结构，而传统的挖掘算法一般仅仅适用于简单的项集。

另外，工业企业内部 DCS 系统的多层分布架构，使得各类监测点的分布呈现多层分布的特点。挖掘各层的关联规则，采用传统的挖掘算法进行各层独立挖掘，算法将产生庞大的候选项集。

1.2 问题分析

针对项集 I 中的各子项具有时间特征的问题，引入时间维度 T，将时间间隔作为筛选条件，对事务集合 D 进行分组。分组需与应用背景结合，以寻求最佳的分组次数。

工业企业内部最小监测点的综合情况由内部的多个方面所决定。对于具有复杂结构的最小监测单元，本文用结构化项集的概念来表示这类项集。结构化项集屏蔽项集的内部结构，使得结构复杂的项集成为一个整体作为挖掘的对象。

对于工业企业内监测点多层分布的特点，引入空间的维度 N，划分挖掘对象的层次，将不同类别的结构监测点转化为 N 层分布结构。

多层分布式项集结构如图 1 所示：

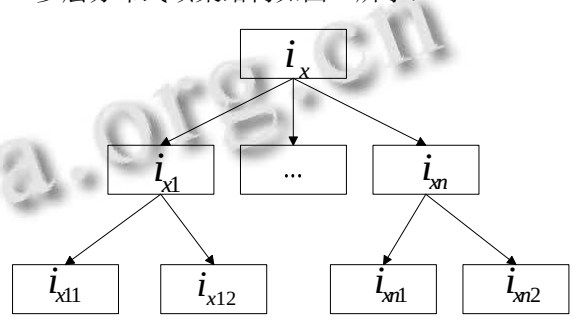

图 1　多层分布式项集结构

引入空间维度后，对结构化项集多层挖掘，采用自顶而下和逐层过滤的挖掘策略，也就是挖掘出高层的项集 i_x 是频繁时，方才考虑 i_x 下一层的子项。采取这样的方式，可有效地节省算法的空间，避免在低层挖掘时产生庞大的候选项集。另外，高概念层项相对于低概念层项而言，成为频繁项集的概率要大于低概念层项集，为此，采用按比例逐层递减的方式设定低层的支持度阀值和自信度阀值，以更好地挖掘低层的关联规则。

2　算法的基本思想与描述

2.1 数据分组

基于以上的分析，当项集 I 中的各项具有时间特征时，本文引入时间维度 T，进行对事务集数据分组。设事务集合 D 为时间跨度 T 内的数据记录。

分组规则：定义 $T=T_1+T_2\ldots+T_n$，以 T_i 为筛选条件，使得：

$$D=\{D_1,D_2,\ldots D_n\}$$

D_i 是 T_i 时间段内的数据集。

设事务集合 D 的最小支持度阈值为 $MinS$，最小置信度阈值为 $MinC$。设定各个数据集 D_i 中最小支持度阈值和最小置信度阈值，可依据以下公式：

$$MinS(D_i)=\frac{S(D_i)}{S(D)}MinS \quad (1)$$

$$MinC(D_i)=\frac{S(D_i)}{S(D)}MinC \quad (2)$$

$S(D_i)$ 为 D_i 的记录总数，用于表示数据集 D_i 的大小。

2.2 结构项集表示

工业企业中监测点的监测结果由内部的多个方面所决定，一方面为了简化挖掘的对象，使得结构不同的项集关系平等，另一方面为了保留最小监测点内部原有的层次结构，以备后续必要的挖掘分析，本文提出用结构化项集表示最小监测点单元，它屏蔽了监测点单元内部子项的结构，使得在同一类别监测点的关系相对单一。

2.3 多层挖掘

工业现场的 DCS 架构，使得不同类别的监测点的分布也呈现出多层分布的特点，为此，引入空间维度，划分挖掘对象的层次。多层结构项集挖掘，在挖掘策略上，采用自顶向下的方式，对每一层的结构项集挖掘，采用经典的 Apriori 算法，在处理层与层之间的关系上，采用逐层过滤的搜索策略，以避免低层产生过多的候选项集。

设从 i 层项集 $I=\{i_1,i_2...i_n\}$ 挖掘出的频繁项集为：$L_k=\{i_x,i_y,i_z\}$。

设结构项集 i_x,i_y,i_z 包含下一概念层次的结构项集，即：$i_x=\{i_{x1},i_{x2},i_{x3},\ldots\}$，$i_y=\{i_{y1},i_{y2},i_{y3},\ldots\}$，$i_z=\{i_{z1},i_{z2},i_{z3},\ldots\}$。

则得到 $i+1$ 层的结构项集为：

$$J=\{i_{x1},i_{x2},i_{x3},\ldots i_{y1},i_{y2},i_{y3},\ldots i_{z1},i_{z2},i_{z3},\ldots\}$$

采用逐层过滤的方式时，$i+1$ 层的最小支持度阈值和最小置信度阈值计算，依据以下公式：

$$MinS_{i+1}=\frac{L(L_k)}{L(J)}MinS_i \quad (3)$$

$$MinC_{i+1}=\frac{L(L_k)}{L(J)}MinC_i \quad (4)$$

L 代表项集的长度。

2.4 算法描述

输入：事务集 D，时间维度 T、时间参数 T_i 或时间分割次数 m（均等分割）、空间层次 N 以及支持度阀值 $MinS$ 和自信度阀值 $MinC$。

输出：多层频繁项集 $L_{i|k}$。

1) 数据分组及得到各数据区的最小支持度阀值：

foreach $T_i \in T$

{ 得到 $D_i \in D$， $MinS(D_i)$ }

2) 输出各个数据区各层的频繁项集：

```
foreach D_i ∈ D
{
  for(i = 0; i < N; i++)
  { L_i1 = 查找第 i 层的项集 I_i 频繁1_ 项集;
    for(k = 2; L_i|k != ∅; k++)
    { C_i|k 由 L_i|k-1 生成的候补集合;
      //计算 t 在数据集合 D_i 中出现的次数
      foreach t ∈ C_i|k
      { if (出现计数小于 MinS_i)
        从 C_i|k 中剔除; }
      L_i|k = C_i|k ;                    }
    return L_i|k ;
    if(i!=N-1)
    { 获得 I_i+1 及 MinS_i+1; }
  }
}
```

3 算法的应用与实现

3.1 应用背景

改进算法以某一大型印染企业作为试点，该企业的工业生产综合能耗主要包括电、水和蒸汽。企业的大型生产车间配置的监测点，根据监测级别分类，分为总监测点和分监测点，总监测点用于监测一个工区的综合能耗，分监测点用于监测一台设备的综合能耗。

总监测点和分监测点的物理结构类似，为此，用最小能耗监测点结构模型来表示一个监测点的物理模型。最小能耗监测点结构模型如图 2 所示：

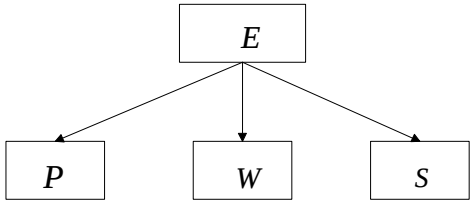

图 2 最小能耗监测点结构模型

E 表示监测点综合能耗情况，P 表示监测点下的电消耗量，W 表示水消耗量，S 表示蒸汽消耗量。

另外，该印染企业在一年中采集了大量的监测数据，这些监测数据跟印染的作业有着重要的相关性，印染的作业是季节性的作业，季节对印染能耗影响程度具有某种相关性，不同季节的产品浸泡在染缸时间将会有明显的差异，这种差异对于不同季节的能源消耗量的影响是不同的。

3.2 具体操作

本文提出应用算法的具体操作如下：

1) 数据分组：将能耗数据库以月份为筛选条件分为 4 个数据区 D_1， D_2， D_3， D_4。

表 1　数据分组

	筛选条件（月份）	数据区
能耗数据	12、1、2	D_1
	3、4、5	D_2
	6、7、8	D_3
	9、10、11	D_4

2) 概念分层：根据不同类别监测点的分布结构，引入空间维度，将试点企业内部各类最小能耗监测结构单元，进行逻辑转化和分层归类，转化为多层架构。

以总监测点 E_i 、分监测点 E_{ij} 及其内部的子项为挖掘对象，构建三层监测结构如下所示：

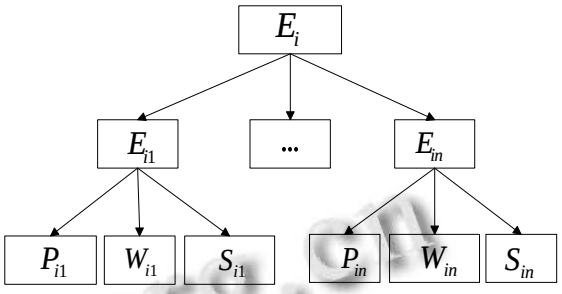

图 3　三层监测结构

3) 输出各层频繁集：采用逐层过滤的方式输出各层的频繁项集。

在数据区 D_1 中，监测数据与各工区能耗总监测点的指标量对比后，转化为事务数据表，某天的部分样本数据如下表所示：

表 2　某天能耗总监测点的预警情况

时间	TID	预警监测点	时间	TID	预警监测点
08:00	1	E_1, E_2, E_3	12:30	10	E_1, E_3, E_9
08:30	2	E_1, E_8	13:00	11	E_2, E_3
09:00	3	E_3, E_5	13:30	12	E_2, E_5, E_7
09:30	4	E_3, E_5, E_8	14:00	13	E_3, E_4, E_9
10:00	5	E_2, E_4, E_8, E_9	14:30	14	E_3, E_5
10:30	6	E_2, E_3, E_7, E_9	15:00	15	E_4, E_6, E_7, E_8
11:00	7	E_2, E_6	15:30	16	E_3, E_5, E_8, E_9
11:30	8	E_2, E_3, E_8	16:00	17	E_3, E_8
12:00	9	E_2, E_7, E_8	16:30	18	E_2, E_4, E_6

设定第一层支持度阀值 $MinS_1 = 22.5\%$，第一层频繁项集输出为：

$$L_{1|3} = \{E_2, E_3, E_8\}$$

设：

$$E_2 = \{E_{21}, E_{22}, E_{23}, E_{24}, E_{25}\}$$
$$E_3 = \{E_{31}, E_{32}, E_{33}, E_{34}\}$$
$$E_8 = \{E_{81}, E_{82}, E_{83}\}$$

则第二层项集为：

$$I_2 = \{E_{21}, E_{22}, E_{23}, E_{24}, E_{25}, E_{31}, E_{32}, E_{33}, E_{34}, E_{81}, E_{82}, E_{83}\}$$

根据 2.3 中公式(3),计算得第二层支持度阀值：

$$MinS_2 = \frac{3}{12} MinS_1 = \frac{1}{4} MinS_1 = 5.63\%$$

则第二层频繁项集输出为：

$$L_{2|4} = \{E_{22}, E_{32}, E_{34}, E_{82}\}$$

设分监测点 $E_{22}, E_{32}, E_{34}, E_{82}$ 的内部子项如下：

$$E_{22} = \{P_{22}, W_{22}, S_{22}\}$$
$$E_{32} = \{P_{32}\}$$

$$E_{34}=\{P_{34},W_{34},S_{34}\}$$

$$E_{82}=\{P_{82},S_{82}\}$$

计算第三层支持度阀值，得：

$$MinS_2=\frac{4}{9}MinS_1=2.50\%$$

则第三层输出频繁项集为：

$$L_{3|5}=\{W_{22},S_{22},P_{32},P_{34},P_{82}\}$$

3.3 结果分析与比对

1) 引入时间维度对频繁项集输出的影响。

样本数据中，由 4.2 的具体操作得到，各个数据组的频繁项集输出如表 3 所示：

表 3 各数据组频繁项集输出

数据区	第一层频繁项集	第二层频繁项集
D_1	E_2,E_3,E_8	$E_{22},E_{32},E_{34},E_{82}$
D_2	E_2,E_4	E_{22},E_{41},E_{42}
D_3	E_2,E_3,E_6	$E_{22},E_{32},E_{34},E_{61}$
D_4	E_2,E_8	E_{22},E_{82}

由表 3 可知，不同的时间段内，输出的频繁项集并不同，时间维度的引入，挖掘出了各个时间段内特定的关联规则，这些关联规则可在各个时间段内灵活应用，以帮助工业企业更准确地掌握能耗预警的规律。

2) 引入时间维度和空间维度对挖掘效率的影响。

引入时间维度对事务集进行分割，引入空间维度，对各类监测点进行层次划分。不同挖掘算法，增加分割次数，对样本数据进行挖掘的实验结果如图 4 所示。

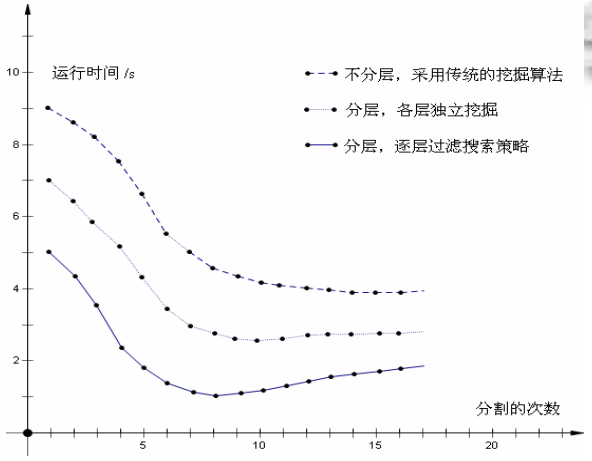

图 4 不同挖掘算法，增加分割次数对执行时间的影响

由图 4 可知，引入时间的维度和空间的维度，采用逐层过滤的挖掘策略进行关联挖掘时，挖掘的效率最高。另外，增加事务集的分割次数，每个数据组事务量减少的同时，各个数据组的事务量也相应地增加，算法运行效率并非一定随着分割次数增加而提高，因此，根据样本数据的实际情况，寻求最佳的分割次数，才能最大程度地提升挖掘的效率。

3) 逐层过滤搜索策略对候选项集存储空间的影响。

引入空间维度，对各类监测点进行层次划分后，设任意第 $i+1$ 层项集 I_{i+1} 的实际长度与第 i 层项集 I_i 之比都为 $m:1$，第一层项集 I_1 的长度为 n。采用传统的 Apriori 挖掘算法，各层独立挖掘，对第 i 层的所有项集进行挖掘，产生的候选项集空间复杂度的上限为 $S(n)$，则：

$$S(n)=2^{nm^{i-1}}-1 \tag{5}$$

采用逐层过滤搜索后，设第 $i+1$ 层项集 I'_{i+1} 的长度与第 i 层项集 I'_i 之比为 $c_i:1$，对第 i 层挖掘时，产生的候选项集空间复杂度的理论上限为 $S'(n)$，则：

$$S'(n)=2^{c_1c_2\cdots c_{i-1}n}-1 \tag{6}$$

设 $c=c_1c_2....c_{i-1}$，因为逐层过滤时，支持度阀值按比例逐层递减，所以在理想情况下，c 接近于 1。

所以，再根据公式(5)和公式(6)，候选项集空间复杂度的上限之比为：

$$\frac{S(n)}{S'(n)}=\frac{2^{nm^{i-1}}-1}{2^{cn}-1}\approx 2^{(m^{i-1}-c)n} \tag{7}$$

越往下层挖掘，项集 I 也将越多，产生候选项集也越为庞大，当设置的支持度阀值较低时，候选项集将占用庞大的存储空间，而采用逐层过滤搜索策略，能够极大地节省算法的空间。

4 结束

本文针对工业企业综合能耗具有的时间特征和空间特征提出一种新的挖掘算法，该算法在经典的 Apriori 算法基础上进行了改进，引入时间维度和空间维度，能够在各个时间段和不同的空间层次中挖掘更多的关联知识，以更好地掌握能耗预警的规律。新算法与传统挖掘算法相比，在算法运行效率上得到了明显的改善。另外，新算法采用逐层过滤的挖掘策略避免了在低层挖掘时产生庞大的候选项集。这种新的时

(下转第 105 页)

及难以确保最后评价结果的准确性和可靠性的现状，根据我国煤矿的实际情况，归纳煤矿事故发生的危险因素和影响矿井生产的不安全因素，最终确定了 9 类 54 个主要影响因素，采用 RBF 神经网络建立煤矿综合安全评价模型，同时为了克服神经网络易陷入局部最小，采用了量子遗传算法对神经网络模型的权值（阈值）进行优化。仿真结果显示，该模型能有效识别样本，并将该方法应用在阜新矿业集团公司某矿，用历史数据进行验证，结果表明，该模型可以准确地评价煤矿安全生产，为煤矿安全评价提供一种有效可靠的方法。

参考文献

1 董建美.我国煤矿事故多发的原因分析及对策.国土资源,2007,(1):22−25.

2 何学秋,等.安全工程学.徐州:中国矿业大学出版社,2000.155−179.

3 国家安全生产监督管理局.安全评价.北京:煤炭工业出版社,2002.1−313.

4 韩斌君.我国煤矿安全事故致因研究[硕士学位论文].上海:同济大学,2007.

5 刘海波,施式亮,刘宝探.人工神经网络对矿山安全状态的评判能力分析.安全与环境学报,2004,(5):69−72.

6 张德丰.神经网络应用设计.北京:机械工业出版社,2009.70−78.

7 王小平,曹立明.遗传算法—理论应用与软件实现.西安:西安交通大学出版社,2002.80−95.

8 Yang JA, Li B, Zhuang ZQ. Multi-universe parallel quantum genetic algorithm its application to blind source separation. Proc. of IEEE Int'l Conf. on Neural Networks & Signal Processing. New York: IEEE Press, 2003,12: 393−398.

9 王凌,吴昊,唐芳.混合量子遗传算法及其性能分析.控制与决策,2005,20(2):156−158.

10 李欣,程春田.曾筠.基于改进量子遗传算法的过程神经元网络训练.控制与决策,2009,24(3):347−351.

(上接第 73 页)

空挖掘算法，同样可以应用到其它具有时间特征和空间特征的工业领域中。

参考文献

1 刘美.关联规则在企业电耗数据分析中的应用.微计算机信息,2009,33:55−57.

2 Fayyad UM, Piatetsky-Shapiro G, Smyth P, et al. Advances in Knoledge Discovery and data mining. AAAI/MIT Press. 1996.

3 刘书暖,田锡天,张振明,许建新,朱名铨.基于 Apriori 算法的典型工序序列获取方法.计算机集成制造系统,2006,12(8):1279−1283.

4 王斌,谢庆生.基于改进遗传算法的制造资源关联规则挖掘.计算机集成制造系统,2007,13(6):1153−1157.

5 沈国强,覃征,沈云斐.一种高效的多维多层关联规则挖掘算法.计算机工程与应用,2008,(12):174−176.

6 李立羽,施鹏飞.OLAP 关联规则挖掘.计算机工程与应用,2002,16(1):128−130.

7 王秋华,王越,曹长修.并行关联规则的挖掘算法的研究.计算机工程,2006,26(2):373−375.

8 李绪成,王保包.挖掘关联规则中 Apriori 算法的一种改进.计算机工程,2002,28(7):104−106.

9 庄晓毅,张忠能.一种改进的关联规则挖掘算法.计算机工程,2004,30(14):128−129.

10 钱光超,贾瑞玉,张然,李龙澍.Apriori 算法的一种优化方法.计算机工程,2008,23(34):196−198.