数据挖掘技术在能量管理系统中的应用

张柏礼[1,2] 孙志挥[1]
1.东南大学计算机科学与工程系 (210096)
2.电力自动化研究院国家电力公司(210003)

Abstract

Large amounts of time-series data are accumulated by energy management system (EMS) database of power system. If we can utilize efficiently these abundant data resource, and adopt the suitable technique of data mining,perhaps we can discover hidden rules and new knowledge which would be great beneficial for the system running more safely and efficiently.

Keywords:data mining,EMS,power system

摘 要

长期运行的电力系统能量管理系统数据库积聚了大量的时态数据,有效地利用这些丰富的数据资源,并采用适当的数据挖掘技术进行处理,可以发现潜在的运行规律和知识,将有助于系统更高效更安全的运行。

关键字:数据挖掘,能量管理系统,电力系统

本文讨论以能量管理系统(EMS)的历史数据库为对象,进行数据挖掘研究和开发的过程。以往对历史数据库中存有的大量时态数据因缺乏有力工具，一般仅用于报表、曲线显示及故障反演，侧重于纵向地考虑,缺少多方位系统性分析;因此对一些事故(包括潮流震荡、频率波动)的成因及发生前后序列,主要依靠有关人员的经验和直觉来决断，未能借助对已有的运行数据数学分析来获得结论和参考，因而影响了事故分析和措施采取的及时性和精确度，更无法发现平时运行中的隐患和存在的规则。数据挖掘是一种从大量数据中智能地、自动地抽取有价值的知识或信息的技术和工具,它已成功地运用于许多领域。同样如将其运用于 EMS 的历史数据库，就可能从中挖掘出先前未知的、有效的、可实用的信息供决策。

1 基本思想

电网运行过程中,各数据间往往是相互制约、相互影响的。例如电压与频率的变化，一般相关于负荷有功、无功或发电机组出力、转角的变化。下面我们以一个简单的实例来说明在关联数据挖掘模型和序列挖掘模型的基础上，进一步根据实际需要对算法进行了适当调整，期望能够减少对数据库的 I/O 扫描次数,并能够适用于 EMS 的需要,发现其特有的关联规则。

以 A、B、C、D、E、F、G、H 八个量取值“0”或“1”代表八个实测量如电压、有功、无功的正常或越限状态,这样可以用一个字节 8 位代表八个量某一时刻的状态,其中 A 值越限或正常对应于第 0 位的值的“1”和“0”,同样 B 值对应了第 1 位的值的“1”或“0”…… H 值对应了第 7 位的值的“1”或“0”。而“code”表示该字节所对应的数值,用一个一维数组 count [code]表示本字节取值为 code 的出现次数。code=0 表示无越限事件发生,code=0×01,0×02,0×04……表示只有一个量越限,code=0×03,0×05,0×06……表示只有二个量越限,即 code 中出现“1”的个数也就是数据项的维数,如 count [0×03]=8 表示 A、B 事件同时发生 8 次,再用 next [code_x] [code_y]表示事件状态为 code_x 的直接后继变为 code_y 的出现频度。

若以一天的 code 变化序列为基本序列单元,code 的采样周期与时态库的保存周期相同为 5 分钟,(对突发事件以秒为单位，但研究的序列不是以天为单位,而以 10 秒钟为单位。所以研究方法和数据量相似),code 的最短变化序列长度为 0,即一天内无事件发生;最长变化序列长度为 288。我们用另一个二维组 record [365] [288+2]来记录一年内所有的 code 变化序列。

2 算法

下面是算法实现的步骤:

第一步:扫描历史库,根据所研究的八个量的越界状态填写数组 record [365] [288+2] (初值为 0),其中 code 发生变化才计入序列,则 record[i]所指以 0×00,0×00 为结尾的串代表了第 i 天的变化序列。

这一步的目的主要将多遍的数据库扫描的 I/O 开销变为一遍，当然是以增加 CPU 和 RAM 开销为代价。

第二步：根据 record 值填写 count [code]（初值为 0）

```
for( I=0;I < 365 ;I ++)
{    for( j=0;j<288;j++)
     {    if (record[I][j]==0&&record[I][j+1]==0)break;
          count[record[I][j]]++;
     }
}
```

第三步：根据 count[]值统计频繁数据项集

```
     for (code=1;code<256;code++)
     {    event_num=EventNum(code);
//计算 code 中"1" 的个数即并发事件数
          switch(event_num)
          {    case 1:                    //一维最小置信度
                    if (count[code]>min_conf[1])
                    {将 code 输入至 C1}
                    break;
               case 2:                    //二维最小置信度
                    if (count[code]>min_conf[2])
                    {将 code 输入至 C2}
                    break;
                                  ……
               case 8:                    //八维最小置信度
                    if (count[code]>min_conf[8])
                    {将 code 输入至 C8}
                    break;
          }
     }
```

这里统计频繁项集的方法与经典模型不同，其主要目的是统计频发事件，发现存在那些频发事件组合形态及其它们的频发程度，也为研究并发的因果关系统提供参考。

第四步：根据 record[]填写 next[][]找出各事件与后继事件的关联度

```
for( I=0;I < 365 ;I ++)
{    for( j=0;j<288;j++)
     {    if (record[I][j]==0&&record[I][j+1]==0)break;
          next[record[I][j]][ record[I][j+1]]++;
     }
}
next[][]实质是建立了一个各事件状态间的转移次数的矩阵
for( I=0;I < 256 ;I ++)
{    for( j=0;j<256;j++)
     {    if (I==j)continue;                    //code_x=code_y
          if (count[I]<min_conf[EventNum[I]])continue;
                                        //code_x 非频繁项不计
          if (next[I][j]/count[I] > min_relation_conf)
//最小因果关联度
               {将 <I,j> 输入至"T" 集中}
     }
}
```

next[][]与 count[]之比也就是由一频繁事件状态转为另一事件状态的发生频度，也就是表现出其间的因果关系的可能性。

3 实验与结论

对于一个区域性的电网，其关口电压、频率、有功、无功并不太多，所以事件的维数在可控范围（几十~几百），加之对小概率事件可以忽略（count、next 较小时作 0 处理），next[][]变成一个很大的稀疏矩阵，可以利用一些矩阵算法进行大规模约解，则在计算开销上可得到有效控制，不失为一种可实用的有效算法。

研究中以梅州地区 EMS 作了一个原型系统的实验验证，该系统采用 SYBASE 数据库，ALPHA DS20 作服务器，已积累近三年的数据量。从中选取较有代表性和较活跃的十六个遥测量为对象，研究它们之间可能的相关性。用一个十六位的短整型量存储这些遥测量状态，选取较为典型的 1999 年数据作为数据空间。考虑到程序的简单和 ALPHA 芯片的高速，实现过程中对 next[][]未采用优化算法，min_conf[1]-min_conf[16] 和 min_relation_conf 取值可在线人工调整，当 min_relation_conf 取为 30%，程序输出七条可能的因果转移规则，当 min_relation_conf 取为 40%，输出四条，其中二条较短时间内得到调度员们的首肯，同时认为并发事件集的输出也有着很好的参考作用。程序的时间开销主要还是集中于第一步对数据库 I/O 的操作，同时为了不影响服务器的正常工作，还需人为减缓 I/O 的频度，因此后几步的时间开销相对而言很小。

以上实验验证了算法的有效性，所得到的并发事件集和因果转移规则，反映了电网中潜在运行规律，而这些规律过去并不为人们所察觉和重视，用这些规律指导电网运行和控制，已为系统的可靠安全运行产生了良好的效益。

参考文献

1 孙志挥，肖利.知识发现与数据挖掘.东南大学计算机科学与工程系，2000

2 王能斌.数据库系统原理.电子工业出版社，2000

3 M.Sforna. Data mining in a power company customer database.Electric Power System Research55,2000：201~209

［收稿日期：2001.11.26］