第 33 卷 第 6 期 计 算 机 工 程 2007 年 3 月
Vol.33 No.6 Computer Engineering March 2007
·博士论文· 文章编号：1000—3428(2007)06—0009—03 文献标识码：A 中图分类号：TP391.77

基于频繁图模式挖掘的质量管理过程分析

陈 亮，高建民，陈 琨
(机械制造系统工程国家重点实验室，西安交通大学 CIMS 所，西安 710049)

摘 要：提出了基于频繁图模式挖掘的工作流频繁活动序列分析的方法，分析了质量管理过程中的关键活动链结构，辅助质量管理过程控制和改进。针对质量管理过程循环结构多且复杂的特点，提出了一种基于Apriori的改进频繁活动序列挖掘算法，设计了新的子图连接算子，减少冗余候选频繁子图的产生。以质量外审意见处理流程的分析为例对该算法进行了说明和分析。

关键词：质量管理过程；过程分析；工作流；频繁图模式挖掘

Analysis of Quality Management Process Based on Frequent Graph Pattern Mining Technology

CHEN Liang, GAO Jianmin, CHEN Kun
(State Key Laboratory for Manufacturing System Engineering, CIMS Institute, Xi'an Jiaotong University, Xi'an 710049)

【Abstract】 Critical activity execution structures are analyzed based on the frequent graph pattern mining technology, supporting the quality management process control and improvement. Aiming to treat with the multiple and complex loop structures in quality management process, this paper puts forward an improved pattern-mining algorithm based on Apriori, in which new sub-graph coalition operator is designed to reduce the generation of redundant frequent candidates. Process analysis of external audit reports treatment is used as a case study to illustrate the proposed approach and its performance in the end of the paper.

【Key words】 Quality management process; Process analysis; Workflow; Frequent graph pattern mining

传统的质量管理理论和方法侧重结果，希望在终端处控制质量。随着ISO9000 族标准的贯彻、深化和TQM的实施，质量控制方向从结果逐渐转向了过程，根据具体过程考虑其资源投入、测量方式和管理方式，这就是所谓的“过程方法”[1]。“过程方法”的提出，对企业质量管理和质量控制提出了新的要求，如何分析以往的过程执行信息，识别关键质量管理过程活动序列模式，指导质量控制和质量改进，是目前工业界和学术界的一大研究热点[2]。过程挖掘是利用数据挖掘技术分析过程日志中存储的过程执行实例信息，辅助过程控制和过程改进工作[3,4]。与传统的数据挖掘对象不同的是，过程一般表现为图的结构，基于频繁图模式的挖掘算法设计是分析过程中频繁活动序列模式的研究重点。

1 工作流频繁活动序列模式挖掘

图作为一种复杂数据存储方式，在实际应用中有着不可替代的作用。下面给出频繁图模式序列挖掘问题的基本定义，并在此基础上引出工作流频繁活动序列挖掘概念。

定义 1 有向图(Direct Graph)。可以描述为一个二元组$G=(V(G),E(G))$，其中$V(G)=\{v_1,v_2,\ldots,v_k\}$是图的顶点集，$E(G)=\{e_h=<v_i,v_j> \mid v_i,v_j\in V(G)\}$是边集。顶点集中，没有输入边的顶点称为源顶点(Source Vertex)，相反，没有输出边的顶点称为沉顶点(Sink Vertex)。有向图中包含的顶点数称为有向图的长度，记为|G|。为了简便起见，如无特殊说明，本文以后提到的“图”均指有向图。

定义 2 子图。给定图$G=(V(G),E(G))$和$G_s=(V(G_s),E(G_s))$，如果满足$V(G_s)\subseteq V(G)$，$E(G_s)\subseteq E(G)$，则称G_s为G的子图，记作$G_s\subseteq G$，也称G包含G_s。

定义 3 图库。图库$GD=\{G_1,G_2,\ldots,G_N\}$，其中，$G_i$为有向图，$N$称为图库的大小。

定义 4 频繁（子）图。给定图库GD，图G的支持度记为$sup(D)$

$$\sup(G)=\frac{|\{G' \mid G\subseteq G'\in GD\}|}{|\{G' \mid G'\in GD\}|}$$

给定最小支持度 minSup，当且仅当$sup(D)\geq$minSup 时，称G为GD中的频繁（子）图。长度为k的频繁图记为k-频繁图。

定义 5 工作流过程模型是业务过程P的规范化描述，是业务活动及其执行方式的定义。过程模型，记为$ws(P)$，可表示为图结构，$ws(P)=(V(G),E(G))$，其中图的节点集V表示过程中的活动集，图的边集E反映过程活动的前序关系。

定义 6 过程实例集，记为$wi(P)$，是工作流过程模型的所有可能的完整执行的集合。对于$\forall I\in wi(P)$，I是包含$ws(P)$的开始节点和结束节点的$ws(P)$的子图。

定义 7 工作流日志，又叫过程日志，简记为$L(P)$，是工作流过程挖掘的图库，记录了业务过程P一段时间内所有实际执行的过程实例产生的活动信息。$L(P)=\{s_1,s_2,\ldots,s_n\}$，其中$s_i$称为过程日志项，是第$i$个过程实例执行中产生的活动信息，可以描述为一个三元组的集合$s_i=\{(v_i,st_i,et_i) \mid v_i\in V\}$，其中$st_i$和$et_i$分别代表$v_i$的开始时刻和结束时刻。考虑到过程执行中可能存在循环，一个活动节点v_i在一条日志项中可能出现多

基金项目：国家自然科学基金资助项目(50505036)；国防科工委支持项目
作者简介：陈 亮(1977－)，男，博士生，主研方向：工作流管理，质量管理；高建民，教授、博导；陈 琨，讲师
收稿日期：2006-04-01 **E-mail：**bluemoon@mailst.xjtu.edu.cn

次。

定义 8 给定工作流日志 $L(P)$和最小支持度 minsup，工作流频繁活动序列模式挖掘的目的是找出 $L(P)$中的所有频繁子图。

2 基于工作流频繁活动序列模式挖掘技术的质量管理过程

质量管理过程网是围绕产品质量，以满足顾客需求为核心，通过对产品质量信息流的分析，将形成和影响产品质量的全部过程连接成一个整体的质量功能网络结构，以期获得最佳产品质量。随着网络化制造和虚拟企业联盟等现代制造技术的发展，质量管理过程越来越复杂，基于工作流技术的协同化质量管理的研究，是处理复杂、动态、多种异构介质共存的质量管理行为的有效方法。

质量工作流系统经过一段时间的执行，工作流平台记录了大量执行实例的活动信息，挖掘分析这些过程实例中频繁出现的活动执行序列模式，对于质量管理过程的分析、控制和改进有不可忽视的作用。

在质量管理过程网络中，PDCA 循环是质量管理过程及质量管理活动应遵守的准则，每一个质量管理过程都应按 PDCA 循环实施闭环管理，周而复始，使产品质量不断改进，相应的，其工作流模型包含了大量结构复杂的循环结构体，一个过程实例执行中，同一活动可能出现多次，从而造成过程日志图库中节点数大量增加，频繁活动序列挖掘过程复杂。本文采用一种基于 Apriori 的算法解决质量管理过程频繁活动序列模式挖掘问题：先构造只含有一个顶点的频繁子图，然后以此为基础，构造含有两个顶点的候选频繁子图集，并通过扫描图库判断候选频繁子图是否是频繁子图，以此类推，从 k-频繁子图集通过连接操作得到候选(k+1)-频繁子图集，在此基础上扫描图库，找出(k+1)-频繁子图集。

连接操作是生成候选频繁集的核心，传统基于 Apriori 的频繁图模式挖掘算法认为：当两个(k-1)-频繁图模式中仅有一个顶点不同，且相同顶点间的所有前序关系相同时，这两个图模式能够连接成为一个候选 k-频繁模式图。这种连接方式将产生大量的冗余候选频繁图模式，为了减少产生的冗余候选频繁图模式，本文提出一种新的连接算法。

定义 9 图的减运算。设 v 是图 G 中的一个顶点，图 G 减去顶点 v，记作 G-{v}，得到的结果是从 G 中删除顶点 v 及其与之相关的边，增加从任意流向 v 的顶点到任意流出 v 的顶点的边所构成的图。

设s是图G的一个源顶点，e是一个沉顶点，由定义 9 不难发现，如果G是频繁的，则G-{s}和G-{e}必将是频繁的，因此，连接两个(k-1)-频繁图模式生成候选k-频繁图模式的时候，只需要分析这两个(k-1)-频繁图的源顶点和终顶点即可，即当 G_i-{s} = G_j-{e}且$s \neq e$时，图G_i和G_j可以连接。

所有的可连接(k-1)-频繁模式两两连接后得到的结果集是候选 k-频繁模式集的超集，根据 Apriori 性质，“一个非频繁(k-1)-项集不可能成为频繁 k-项集的一个子集”，对该结果集进行修剪，就能得到候选 k-频繁模式图。

候选频繁图生成算法过程伪代码描述如下：

```
GenerateCandidate (Lk-1):候选频繁子图集 {
  CandidateSet = Ø;
Forall (<Gi,Gj>| Gi,Gj ∈ Lk-1) {
        Forall (s | (s, t) ∈ E(Gi), t ∈ V(Gi)) {
          Forall (e | (o, e) ∈ E(Gj), o ∈ V(Gj)){
        If (Gi-{s}=Gj-{e}) { //两图模式可连接
            UG1 = Gi∪Gj;
            UG2 = Gi∪Gj ∪ {(s, e)};
            CandidateSet=CandidateSet∪{UG1};
            If ((s,e) ∉ E(UG1))
            CandidateSet=CandidateSet∪{UG2};  }
            }  }  }
    Forall(G∈CandidateSet){//候选频繁模式修剪
      If (DeriveSub (G)∩Lk-1≠DeriveSub (G))
        CandidateSet = CandidateSet – {G};    }
    Return CandidateSet;  }
```

候选频繁模式修剪时，需要找出连接结果集中的 k-频繁候模式图的所有长度为(k-1)的子图，该过程算法伪代码如下：

```
DeriveSub(G: k-图): (k-1)-子图集  {
  Subgraph = Ø;
  Forall (v ∈ V(G)){
    Source = {tv | (tv,v) ∈ E(G), tv ∈ V(G)};
    Sink = {fv | (v,fv) ∈ E(G), fv ∈ V(G)};
    SG = G – {v};
    Forall ( (vs,vd) | vs ∈ Source,vd ∈ Sink) {
      if ( (vs,vd) ∉ E(SG) )
        SG = SG ∪ {(vs, vd)};      }
  Subgraph = Subgraph ∪ {SG};  }
  Return Subgraph;  }
```

考虑到过程模型可能存在循环，一个活动在一条过程日志项中可能出现多次，处理时，先将同一活动的不同执行实例按照开始时间先后顺序依次标注上脚标 1，2，…，算法处理过程中当作不同的活动节点，处理完后，将同一活动的不同执行实例合并成为过程模型的一个活动节点。

3 算法实例

某飞机制造公司定义的外审意见处理简化过程描述如下：首先，外审方根据年检验计划对公司的质量管理体系以及产品质量过程进行外部审核，发现质量问题后下发不合格项报告；责任单位负责人根据不合格项报告制定纠正措施计划，并将计划分别提交给标质处业务员和技术处业务员进行审查，对于其中某些外购关键件的技术指标，技术处需要与外购方进行协商制订。审核结果汇总提交给纠错主管，如果审查通过，由纠错主管将纠正措施计划提交给外审方，如果审查未通过，发回责任单位重新制定纠正措施计划。为了描述方便，以下将外审方下发不合格项报告、责任单位制定纠正措施计划、标质处审核纠正措施计划、技术处审核纠正措施计划、外购方审核纠正措施计划、纠错主管验收纠正措施计划和提交纠正措施计划活动分别记为 A、B、C、D、E、F 和 G。

该过程经过一段时间的执行，工作流平台记录下大量过程日志信息。假设过程日志中仅有两条过程日志项(a)、(b)，其对应的过程实例活动时序图如图 1 所示。

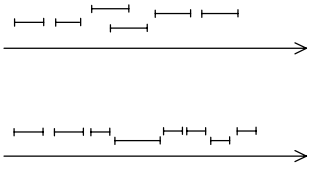

图 1 过程日志中包含的两过程实例活动时序

其中，图 1(a)对应的过程实例满足预定的质量指标要求——包括过程执行周期和质量成本，图 1(b)对应的过程实例的执行周期指标虽然满足，但是质量成本花费过高，因此，根据质量成本指标，可以将过程日志划分为合格过程日志和

不合格过程日志 2 部分，其中，合格过程日志包含过程日志项图 1(a)，不合格过程日志包含过程日志项图 1(b)，分析对比 2 个日志图库中包含的不同频繁活动序列模式，可以得出影响质量成本的关键活动链结构，指导质量过程控制点的选取。

设最小支持度为 100%，限于篇幅，图 2 仅描述了合格过程日志图库与不合格过程日志图库中包含 B、C、D、F 4 个活动的频繁活动序列结构生成过程。

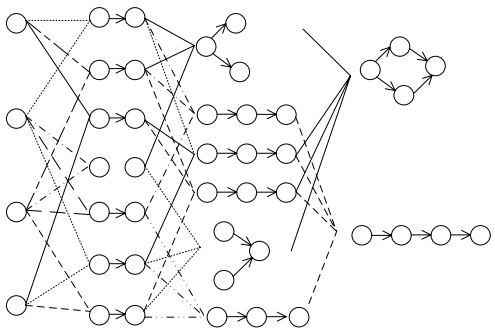

图 2 频繁活动序列结构生成过程片段

对比合格与不合格过程日志图库中蕴含的频繁子图结构可知：C、D 活动的执行次序，即过程中标质处与技术处对纠正措施计划的审核活动采取串行还是并行执行，是影响整个外审意见处理过程性能指标的一个关键活动链结构体。同样，采用该方法还可以得出，技术处与外购方之间的交互循环也是一个关键质量活动链结构。这些关键活动链结构是选取质量管理过程控制点的依据。

4 结论

与传统的频繁序列模式挖掘不同的是，过程一般表现为图的形式，且由于质量管理过程的特殊性，其过程图中含有大量关系复杂的循环子结构，造成过程执行中活动数迅速增加。本文提出一种基于 Apriori 的图模式挖掘算法，设计了一种适合质量工作流子图模式的连接操作算子，降低算法的时间复杂度，提高算法性能。实例证明，该算法能够较好地分析质量管理过程中的频繁活动序列模式，指导质量管理过程控制和改进。

参考文献

1 杨志坚, 丁伯坚, 丁炳山. 2000 版 ISO9000 国际标准术语手册[M]. 北京: 国防工业出版社, 2004.

2 Manoochehr N, Dennis F K. An Integrated Framework for Post-ISO 9000 Quality Development[J]. International Journal of Quality & Reliability Management, 2000, 17(3): 226-258.

3 Geppert A, Tombros D. Logging and Post-mortem Analysis of Workflow Executions Based on Event Histories[C]//Proceedings of the 3rd International Conference on Rules in Database Systems. 1997: 67-82.

4 Casati C, Fabio C, Umeshwar D, et al. A Comprehensive and Automated Approach to Intelligent Business Process Execution Analysis[J]. Distributed and Parallel Database, 2004, 16 (3): 239-273.

(上接第 8 页)

表 1 SVM 与 BP 算法的比较

	识别算法	训练时间(s)	训练次数
1	附加动量 BP 算法	40.878 8	8 906
2	自适应学习速率 BP 算法	5.337 7	1 309
3	弹性 RPROP 算法	1.642 4	310
4	Fletcher-Reeves 共轭梯度法	0.661 0	30
5	Polak-Ribiere 共轭梯度法	0.671 0	38
6	Powell-Beale 共轭梯度法	0.784 1	51
7	BFGS 拟牛顿法	0.901 3	56
8	LM 优化 BP 算法	0.390 6	5
9	SVM 识别算法	0.020 0	2

经过多次仿真和分析表明，对于车型的识别问题，BP 算法的正确识别率在 80%～90%之间。在所有算法中，LM 算法是最快的，而且迭代的步数最少，其主要缺点就是所需的存储量大，需要存储近似 Hessian 矩阵 JTJ，该矩阵是 n×n 维的，其中 n 是网络中权值和偏置值的总数。当 n 过大时，LM 算法可能不是很实用了。共扼梯度算法通常快于学习速率可变和加入动量项的算法，需要存贮空间要多一些，适用于有大量权值的网络。实验结果表明，SVM 识别系统对训练样本的训练时间最短，是 BP 算法中最快的 LM 算法的 13 倍，识别的正确率远远高出 BP 神经网络。

4 结束语

应用改进的双帧差“或”运算法，能够实时快速地检测出运动车辆，并定位出具体位置，测量出车型数据，再根据 SVM 算法对车辆类型进行识别分类，通过与 BP 神经网络识别系统相比较表明，在训练样本较少的情况下，该系统的识别率高于 BP 神经网络，并具有算法简单、无需先验知识、容易控制和稳定性好等优点，可广泛地应用到智能交通信息系统中。

参考文献

1 Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer Verlag, 1995.

2 Collins R, Lipton A, Kanada T, et al. A System for Video Surveillance and Monitoring[R]. Pittsburgh: Robotics Institute, Carnegie Mellon University, 2000.

3 Cucchiara R, Grana C, Piccardi M. Improving Shadow Suppression in Moving Object Detection with HSV Color Information[C]//Proc. of IEEE Int'l Conference on Intelligent Transportation Systems. 2001: 334-339.

4 Platt J C. Fast Training of Support Vector Machine Using Sequential Minimal Optimization[M]. Cambridge, MA: MIT Press, 1999: 185-208.

5 Sebald D J, Buchlew J A. Support Vector Machines and the Multiple Hypothesis Test Problem[J]. IEEE Trans. on Signal Processing, 2001, 11(49): 2865-2872.

6 Chapelle O, Vapnik V, Bacsquest O, et al. Choosing Multiple Parameters for Support Vector Machines[J]. Machine Learning, 2002, 46(1): 131-159.