1) 讨论算法的时间复杂度

由于现有的文献[2,10]已经很好说明了 Alpha 算法可以很好地在多项式时间内容挖掘出过程模型,本文是在此基础上进行了扩展,所以只给出扩展部分算法的时间复杂度分析.令 W 为满足局部完备性的日志,其中轨迹数量为 t,总任务数量为 n,日志事件总数量为 m,则日志抽象中两个主要步骤建立基本次序矩阵和循环比较次序向量,其时间复杂分别为 $O(m)$和 $O(n^2)$;事件特征挖掘中由于两种类型循环可以并行执行,所以仅考虑其中三角形 2 度循环结构的时间复杂度为 $O((n/2)\times m)$;并发分支结构干扰中采用紧邻度来解决的识别混乱问题时间复杂度为 $O((m/2)\times t)$;基于次序向量,采用广度优先搜索算法进行长循环抽象及还原,时间复杂度为 $O(n)$;扩展经典的 Alpha 算法中,还原变体结构的时间复杂度为 $O(n)$.由于α^{L+}算法是顺序执行上面步骤内容,所以整体复杂度为 $O(m+n^2+(n+t)\times m/2+n)$.更为重要的一点是:相对人工建模的业务过程时间(几周~月),采用过程挖掘算法从日志中得到的模型时间(几分钟~几十分钟)可以忽略不计.如图 9 的 W_5 日志在个人 PC 上的挖掘时间为 1ms,而针对较为复杂的实际日志(Log of Volvo IT incident management system,包含 7 554 条轨迹,65 533 个事件,来源 http://data.4tu.nl/repository/uuid:500573e6-accc-4b0c-9576-aa5468b10cee)在个人 PC 上的挖掘时间也只有十几分钟.因此,时间一般不是挖掘算法的瓶颈问题.

2) 讨论现有循环挖掘算法与本文算法的对比

首先给出过程模型中具有 2 度循环的挖掘准确率对比柱状图(图 10 所示),充分展示了本文算法相对原有算法的优势;其次,将现有 2 度循环挖掘的算法集成为一种,统称为集成算法,再与本文算法进行对比(图 11 所示),表明本文算法依然具有优势.

图 10 中,横坐标轴代表的是过程模型中任务的数量,并且长循环及 2 度循环的数量占任务数量的 10%;纵轴代表不同算法对应过程模型随机产生的局部完备性日志挖掘出循环的准确率.每种类型的模型,随机产生 200 个日志文件,然后依次采用现有算法进行挖掘.增量式挖掘算法针对 2 度循环挖掘能力最弱,基本上只能解决长循环结构,对于部分没有处于并发分支上的 2 度循环也能处理;Alpha+挖掘算法强制规定了日志轨迹必须同时存在“*aba*”与“*bab*”模式来满足循环完备性,以此挖掘 2 度循环,而模型复杂度增加会导致同时出现的概率大幅度下降;启发式是目前较为认可的方式来挖掘 2 度循环,即通过“*aba*”或“*bab*”出现的频率,设定一个阀值,满足阀值要求即可.其思想核心也是因为 2 度循环结构体现在日志中的事件特征是“*aba*”模式;本文算法α^{L+}在此基础上更进一步,考虑了事件特征不仅体现在紧邻模式上,更表现为事件间的位置及次数.正常分析,本文算法应该是随着模型规模增长,其准确率与其他算法会逐渐拉开差距,但是在规模为 100 个任务时,本文算法与启发算法基本上是一样的准确率.其主要原因是,紧邻度还不能完全保证百分百地排除识别混乱的问题;其次,模型产生的日志不含有“*aba*”模式的概率虽然是随着模型复杂度增加而降低,但并不是完全成正比关系.不过,从整体上分析,本文算法相对现有算法还是能够较好地解决 2 度循环挖掘问题.

3) 进一步讨论本文算法的内容

本文α^{L+}算法可以从含有或不含有“*aba*”模式的日志轨迹中挖掘出 2 度循环,但是还是不能做到百分之百的准确率.为此,针对现有的实验数据,逐步执行算法的每部分内容,分析当前算法每部分的贡献度及不足.贡献度是为了刻画α^{L+}算法在挖掘 2 度循环过程中,每部分核心内容对挖掘出最终 2 度循环的重要性.贡献度的计算为

$$CT=\frac{SW'}{SW}.$$

借鉴文献[18]用并行度来衡量挖掘后模型的效率问题,本文使用贡献度来衡量算法核心内容的重要性.其中,CT 为贡献度,分子 SW'表示算法核心内容单独或者与其他内容联合执行,能够挖掘出来 2 度循环的日志数量.而分母 SW 表示所有内容联合起来,挖掘出 2 度循环的日志的数量.

如图 11(a)所示,随着模型规模的增加,事件特征内容呈递减下降,而其他 3 部分内容呈缓慢递增上升.其反映了本文算法是以事件特征为核心,其他内容为辅的挖掘方法.在模型规模较少时,紧靠事件特征基本上就可以挖掘出 2 度循环;但随着模型的复杂度增加,变体结构、嵌套结构的出现,需要对日志进行预处理,才能保证日志轨迹也能很好地体现出 2 度循环的事件特征.图 11(b)针对 300 规模模型产生的数据,具体给出了算法中不同内容

的贡献度柱状图:除了 3 层以上的嵌套结构不进行探索以外,目前主要核心内容的不足体现在紧邻度.虽然紧邻度能够较好地解决识别混乱问题,但是其核心依据是真正2度循环的任务间紧邻度要比其他假的2度循环结的任务强.这种解决方式只是根据实际实验判断而形成的,存在一定的小概率事件.即存在识别混乱问题,依靠紧邻度还是不能解决.此外,多重嵌套结构也会造成变体抽象和回路抽象出现错误,从而导致挖掘的2度循环缺失.这部分内容已经在上述讨论中给出,这里不再赘述.综上所述,后续研究要具体针对这些不足进行深入探讨.

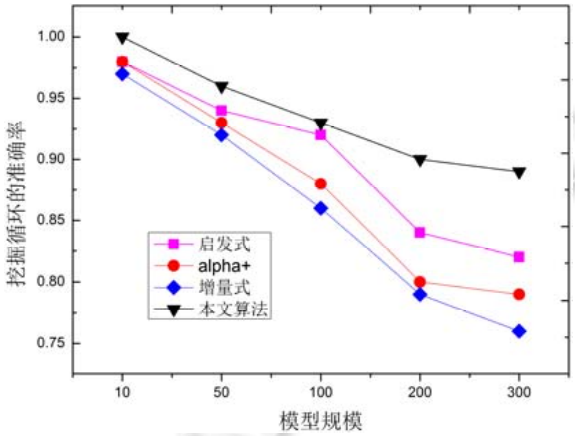
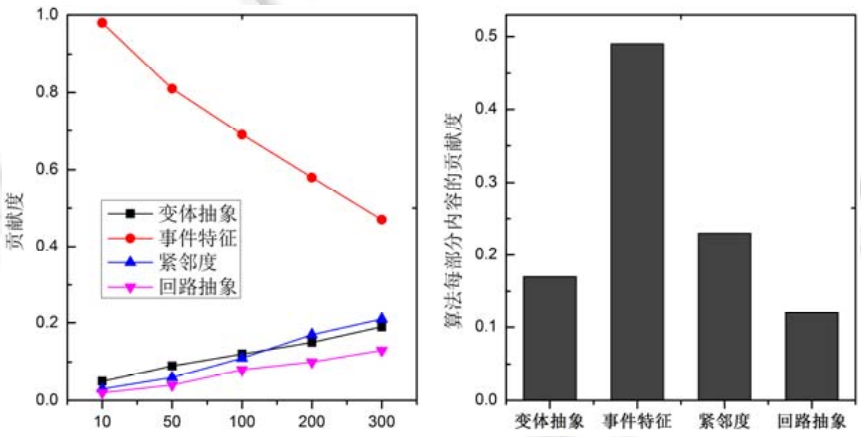


Fig.10 Comparison of precision among different algorithms

图 10 不同算法准确率对比



(a) 模型规模 (b) α^{L+}算法内容

Fig.11 Contribution of each part in α^{L+} algorithm

图 11 α^{L+}算法每部分的贡献度

5 相关工作

过程挖掘已经成为现代组织用于管理复杂运作流程的一种重要手段,其中,准确地挖掘出过程模型是该领域极其重要的研究内容.现有的算法在挖掘模型中的顺序关系、选择关系、并发关系及长循环关系中都获得了较好的进展.最大的难题是如何从日志文件中挖掘出长度为 2 的短循环结构.该内容的讨论及研究也从未停止过,表 4 给出了在挖掘循环结构上取得一定成果的文献内容(按时间降序).

Table 4 Researches about mining loop

表 4 循环挖掘的相关研究

文献	解决问题	技术方法	方式类别
文献[19]	短循环及长距离依赖	运用启发式定义循环因子	“*aba*”模式
文献[20]	不可见任务、循环和非自由网	遗传算法+启发式算法	“*aba*”模式
文献[21]	长循环及任务簇之间依赖关系	定义循环四元组,采用组合案例簇	无
文献[16]	长循环及部分简单短循环结构	增量式挖掘	无
文献[22]	短循环及噪音干扰问题	根据阀值,设定采用启发式挖掘算法	“*aba*”模式
文献[23]	长循环及部分并发干扰结构	定义四元组,采用启发式扩展α算法	无
文献[3]	短循环	启发式挖掘算法	“*aba*”模式
文献[10]	短循环	定义循环完备性的日志	“*aba*”模式

表 4 展示了循环结构挖掘的一些代表性成果,其中,短循环特指本文研究的 2 度循环结构,方式为模式代表其挖掘循环的内容用到了本文前面所提及的“*aba*”模式.从表 4 可以得出,现有的挖掘算法能够很好地解决长循环的挖掘问题[16,21,23].相比之下,短循环的结构挖掘是一个亟待解决的问题.为此,文献[10,19,20]分别提出了启发式因子、遗传+启发、扩展α算法及定义循环完备等方式来解决 2 度循环的挖掘难题.上述解决 2 度循环挖掘问题的本质在于采用了“*aba*”模式,但现实情况是,当日志满足局部完备性时,其日志轨迹中可以不出现“*aba*”模式.因此,本文在前人的工作基础上提出了α^{L^+}算法,用于解决从满足局部完备性且不含“*aba*”模式的日志轨迹中挖掘出 2 度循环.

当前,业务过程随着经济发展而变得相当庞大复杂,要想保证信息系统记录的日志能够完全反映任务间的关系(全局完备性),几乎是不可能的.已有较多的研究工作者提出了在更弱的完备性日志中挖掘出过程模型,如,文献[24,25]分别提出了因果完备性(causally complete)和弱完备性(weakly complete),但此类挖掘算法往往不能得到完整的模型(如并发结构和循环结构不能准确挖掘).因此,本文算法的优势在于不仅放宽日志的约束条件,又能保证准确地挖掘出完整的过程模型.

6 结束语

现有的过程挖掘算法在挖掘 2 度循环时,规定了日志中必须存在“*aba*”模式.但是,满足局部完备性的日志文件可以不存在该模式.为此,本文扩展了 Alpha 算法,从不具备“*aba*”模式的日志中挖掘出 2 度循环结构.本文开始给出了最简 2 度循环结构的定义,并通过次序向量对原始日志进行抽象,排除 2 度循环的变体结构;然后,通过全局视角,采用事件特征来挖掘两种不同的 2 度循环结构及并发结构;接着,针对并发分支上同类型循环结构带来的识别混乱和识别错误问题,提出了紧邻度和回路抽象的概念,较好地解决了以上问题;最后,借鉴前人工作的基础,提出了扩展算法α^{L^+}用于从不具备“*aba*”模式的日志中挖掘包含 2 度循环结构的完整过程模型.此外,运用人工案例及实际案例验证了α^{L^+}算法的可实现性,并通过实验的对比,探讨了目前算法的一些局限性.因此,后续研究也将在本文的基础上,对算法的不足进行更进一步的研究分析.另外,如何在不满足局部完备性的日志中挖掘出合理模型,以进一步提高算法在实际应用中的鲁棒性,也是未来工作的一个重点.

References:

[1] van der Aalst W. Process Mining: Discovery, Conformance and Enhancement of Business Proceses. Springer Publishing Company, Incorporated, 2014.

[2] van der Aalst W, Weijters T, Maruster L. Workflow mining: Discovering process models from event logs. IEEE Trans. on Knowledge & Data Engineering, 2004,16(9):1128−1142.

[3] Weijters AJMM, Aalst WMP, Medeiros AKA. Process mining with the heuristics miner algorithm. Eindhoven University of Technology, 2006,166:1−34.

[4] Medeiros AKAD, Weijters AJMM, Aalst WMPVD. Genetic process mining: An experimental evaluation. Data Mining & Knowledge Discovery, 2007,14(2):245−304.

[5] Sarno R, Sungkono KR. Hidden Markov model for process mining of parallel business processes. Int'l Review on Computers & Software, 2016,11(4):290−306.

[6] van der Werf JMEM, Dongen BFV, Hurkens CAJ, *et al*. Process discovery using integer linear programming. Fundamenta Informaticae, 2008,94(3):368−387.

[7] Bergenthum R, Desel J, Lorenz R, *et al*. Process mining based on regions of languages. In: Proc. of the Int'l Conf. on Business Process Management. Springer-Verlag, 2007. 375−383.

[8] Yang HD, Wen LJ, Wang JM. An approach to evaluate the local completeness of an event log. In: Proc. of the 12th IEEE Int'l Conf. on Data Mining (ICDM 2012). 2012. 1164−1169.

[9] Hofstede AHMT. Estimating completeness of event logs. Technical Report, No.04, BPM Center, 2012.

[10] Medeiros AKAD, Dongen BFV, Weijters AJMM. Process mining: Extending the α-algorithm to mine short loops. Eindhoven University of Technology, 2004,133:145−180.

[11] Yuan CY. Petri Net Application. Beijing: Science Press, 2103 (in Chinese).

[12] Günther CW. Activity mining by global trace segmentation. In: Proc. of the Int'l Workshops on Business Process Management Workshops (BPM 2009). Ulm: Revised Papers, 2009. 128−139.

[13] Polyvyanyy A, García-Bañuelos L, Dumas M. Structuring acyclic process models. Information Systems, 2010,37(6):518−538.

[14] Polyvyanyy A, García-Bañuelos L, Fahland D, Weske M. Maximal structuring of acyclic process models. The Computer Journal, 2014,57(1):12−35.

[15] Zhu R, Li T, Mo Q, Dai F, Gao TL, He Y, Sun X. Heuristic parallelized mining single firing sequence. Computer Integrated Manufacturing Systems, 2016,22(2):330−342 (in Chinese with English abstract).

[16] Ma H, Tang Y, Wu LK. Incremental mining of processes with loops. Int'l Journal on Artificial Intelligence Tools, 2011,20(01): 221−235.

[17] Van Dongen BF, De Medeiros AKA, Verbeek HMW, *et al*. The ProM framework: A new era in process mining tool support. In: Proc. of the Int'l Conf. on Applications and Theory of Petri Nets. Springer-Verlag, 2005. 444−454.

[18] Jin T, Wang J, Yang Y, Wen L, Li K. Refactor business process models with maximized parallelism. IEEE Trans. on Services Computing, 2016,9(3):456−468.

[19] Lu FM, Zeng QT, Duan H, Cheng JJ, Bao YX. College of information science and engineering parallelized heuristic process mining algorithm. Ruan Jian Xue Bao/Journal of Software, 2015,26(3):533−549 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4769.htm [doi: 10.13328/j.cnki.jos.004769]

[20] Vázquez-Barreiros B, Mucientes M, Lama M. ProDiGen: Mining complete, precise and minimal structure process models with a genetic algorithm. Information Sciences, 2015,294:315−333.

[21] Lu FM, Zeng QT, Bao YX, Duan H, Zhang H. Mining algorithm of task dependencies based on process case clusters. Computer Integrated Manufacturing Systems, 2013,19(8):1771−1783 (in Chinese with English abstract).

[22] Wang HY. A process mining algorithm for cycle tasks [MS. Thesis]. Tianjin: Hebei University of Technology, 2011 (in Chinese with English abstract).

[23] Wu S. An extended alpha mining algorithm for complex loop structures [MS. Thesis]. Harbin: Harbin Engineering University, 2011 (in Chinese with English abstract).

[24] Lekić J, Milićev D. Discovering block–structured parallel process models from causally complete event logs. Journal of Electrical Engineering, 2016,67(2):111−123.

[25] Lekic J, Milicev D. Discovering models of parallel workflow processes from incomplete event logs. In: Proc. of the Int'l Conf. on Model-Driven Engineering and Software Development. IEEE, 2015. 477−482.

附中文参考文献:

[11] 袁崇义.Petri 网应用.北京:科学出版社,2013.

[15] 朱锐,李彤,莫启,代飞,高提雷,何云,孙雪.启发式并行化单触发序列挖掘算法.计算机集成制造系统,2016,22(2):330−342.

[19] 鲁法明,曾庆田,段华,程久军,包云霞.一种并行化的启发式流程挖掘算法.软件学报,2015,26(3):533−549. http://www.jos.org.cn/1000-9825/4769.htm [doi: 10.13328/j.cnki.jos.004769]

[21] 鲁法明,曾庆田,包云霞,段华,张昊.基于流程案例簇的任务关系挖掘算法.计算机集成制造系统,2013,19(8):1771−1783.

[22] 王海燕.面向循环任务的过程挖掘算法研究[硕士学位论文].天津:河北工业大学,2011.

[23] 吴苏.一种可发现复杂循环结构的扩展α过程挖掘算法[硕士学位论文].哈尔滨:哈尔滨工程大学,2011.

林雷蕾(1989－),男,海南万宁人,硕士,主要研究领域为软件工程,流程管理,系统分析与集成.

周华(1963－),男,博士,研究员,博士生导师,主要研究领域为软件工程,系统分析与集成.

代飞(1982－),男,博士,副教授,CCF 专业会员,主要研究领域为软件工程,业务过程管理.

朱锐(1987－),男,博士,讲师,CCF 专业会员,主要研究领域为软件过程,过程挖掘.

李彤(1963－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为软件工程.