区别于现有的软件过程挖掘方法,本文针对软件过程日志的单实例性特征,提出了双层次的挖掘方法.该方法首先从过程日志中通过聚类发现活动信息,然后以活动发生的时序顺序转化为单触发序列,进而通过循环将单触发序列划分为多个循环实例,这些循环实例能够作为案例信息支持传统过程挖掘方法的使用.

由于通过过程日志发现活动信息后将形成单触发序列,因此本文以单触发序列中的循环为着手点,将单触发序列进行循环实例划分,以达到发现循环部分的案例信息,进而用于过程挖掘.因此,本文方法是对当前过程挖掘方法的完善,在处理单触发序列方面,相对于其他算法具有明显优势.

对于单触发序列的研究可能还与序列挖掘、情节(episode)挖掘以及 Petri 网语言等方面相关,其中,序列挖掘的目的是为了发现序列中的频繁序列[56];情节挖掘[57]是为了发现频繁的情节,一段情节定义了一个偏序关系,其挖掘方法常常使用滑窗的方式进行,情节挖掘不支持发现并发活动,序列挖掘和情节挖掘都是关联规则学习的变体,它们都只考虑顺序关系而不支持发现并发、选择和循环;Petri 网语言[58]主要是指 Petri 网及其所产生语言之间的关系,它主要被用来分析 Petri 网的行为,模型对应语言,语言产生轨迹,通过该过程相对容易,但是从轨迹发现语言(或模型)相对较难.总之,上述方法都不能完全有效地解决单触发序列的挖掘问题,而对单触发序列的解决是本文双层次挖掘方法与其他软件过程挖掘方法最大的区别.

## 8 结束语

当前,软件过程建模问题已经成为限制软件过程研究的核心问题,过程挖掘作为数据科学的一种最佳实践,在多种领域中得到了应用和推广,以二者结合的方法来对软件过程中涉及到的数据和挖掘方法还鲜有文献进行讨论,因此,本文在此背景下提出了双层次的软件过程挖掘方法.该方法将分为活动层及过程层.

- 在活动层对过程日志中的特征词进行抽取;为了能够有效区别某些单词的重要程度,提出了 WSLVM 模型对每条记录的特征进行向量化,将事件转化为特征向量集,然后对特征向量集进行聚类,将聚类的结果作为活动与事件进行绑定;提出了利用模糊聚类方法并结合平均活动熵来对聚类结果进行确定的方法.
- 在过程层,基于启发式的单触发序列挖掘方法,针对非完全循环的情况下的事件日志的完备性问题进行了研究,提出了非完全循环情况下的循环归属条件.

通过两个真实的软件过程数据来分别进行实验,证明各层次方法的正确性及可行性.

未来工作中,我们拟考虑如下若干方面的工作.

(1) 针对当前的过程挖掘算法进行改进,提出适用于软件过程挖掘的针对性算法.软件过程挖掘的目标在于能够快速地发现一个简洁、合理、优质的过程模型以支持软件开发活动的进行.尽管当前已经存在了一些过程挖掘算法,但现实情况是,软件过程相对复杂,挖掘算法效率较低,挖掘结果不够简洁,不能在多种度量属性间获得平衡.

(2) 从数据来源来看,当前仍然没有一种面向过程挖掘与分析的软件过程仓库管理系统,当前软件过程中的数据都是被动式、面向开发者的.所谓被动式是指当前的执行数据是被动式进行存储的,而不是有目的地对事件信息进行存储.因此,有必要对面向过程的软件过程数据管理系统进行研究.

(3) 软件过程挖掘工具.当前,过程挖掘工具都仅限于客户端、专业人士,这极大地限制了过程挖掘方法的普及和推广.当前,过程挖掘方法已经较为成熟,能够利用现有的优秀的前端展示技术对结果进行展示,将过程挖掘的业务逻辑封装成服务以供不同需求对这些业务服务的使用.

**References**:

[1] CMMI PT. CMMI® for development, improving processes for better products. Version 1.2. Pittsburgh: Software Engineering Institute, 2006.

[2] Mordal K, Anquetil N, Laval J, Serebrenik A, Vasilescu B, Ducasse S. Software quality metrics aggregation in industry. Journal of Software Evolution & Process, 2013,25(10):1117−1135. [doi: 10.1002/smr.1558]

[3] Lonchamp J. A structured conceptual and terminological framework for software process engineering. In: Proc. of the 2nd Int'l Conf. on Software Process, Continuous Software Process Improvement. Berlin: IEEE, 1993. 41−53. [doi: 10.1109/SPCON.1993.236823]

[4] Kaur R, Sengupta J. Software process models and analysis on failure of software development projects. Int'l Journal of Scientific and Engineering Research, 2011,2(2):1−4.

[5] Maciel RSP, Gomes RA, Magalhães AP, Silva BC, Queiroz JPB. Supporting model-driven development using a process-centered software engineering environment. Automated Software Engineering, 2013,20(3):427−461. [doi: 10.1007/s10515-013-0124-0]

[6] Kindler E, Rubin V, Schäfer W. Activity mining for discovering software process models. Software Engineering, 2006,79:175−180.

[7] Rubin V, Günther CW, van der Aalst WMP, Dongen EKBFV, Schafer W. Process mining framework for software processes. In: Wang Q, Pfahl D, Raffo DM, eds. Proc. of the Software Process Dynamics and Agility, Vol.4470. Berlin, Heidelberg: Springer-Verlag, 2007. 169−181. [doi: 10.1007/978-3-540-72426-1_15]

[8] Rubin V, Lomazova I, van der Aalst WMP. Agile development with software process mining. In: Proc. of the 2014 Int'l Conf. on Software and System Process. Nanjing: ACM Press, 2014. 70−74. [doi: 10.1145/2600821.2600842]

[9] van der Aalst WMP. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Berlin, Heidelberg: Springer- erlag, 2011. [doi: 10.1007/978-3-642-19345-3]

[10] Hindle A, Godfrey MW, Holt RC. Software process recovery using recovered unified process views. In: Marinescu R, ed. Proc. of the 2010 IEEE Int'l Conf. on Software Maintenance (ICSM). IEEE, 2010. 1−10. [doi: 10.1109/ICSM.2010.5609670]

[11] Carmona J, Cortadella J. Process discovery algorithms using numerical abstract domains. IEEE Trans. on Knowledge & Data Engineering, 2014,26(12):3064−3076. [doi: 10.1109/TKDE.2013.156]

[12] Yu SS, Zhou SG, Guan JH. Software engineering data mining: A survey. Journal of Frontiers of Computer Science & Technology, 2012,6(1):1−31 (in Chinese with English abstract). [doi: 10.3778/j.issn.1673-9418.2012.01.001]

[13] Xie T, Thummalapenta S, Lo D, Liu C. Data mining for software engineering. Computer, 2009,42(42):55−62. [doi: 10.1109/MC.2009.256]

[14] Rodriguez D, Garcia E, Sanchez S, Nuzzi SRS. Defining software process model constraints with rules using OWL and SWRL. Int'l Journal of Software Engineering and Knowledge Engineering, 2010,20(4):533−548. [doi: 10.1142/S0218194010004876]

[15] Li T. An Approach to Modelling Software Evolution Processes. Springer-Verlag, 2008. [doi: 10.1007/978-3-540-79464-6]

[16] He KQ, Li B, Ma Y, Huang Y. The key technology of software engineering in big data age. Communications of the China Computer Federation, 2014,10(3):8−18 (in Chinese with English abstract).

[17] Wang QX, Mei H. Big data and software engineering. Communications of the China Computer Federation, 2014,10(3):6−7 (in Chinese with English abstract).

[18] Van der Aalst WMP, Adriansyah A, De Medeiros AKA, Arcieri F, Baier T, Blickle T, Bose JC, Brand PVD, Brandtjen R, Buijs J. Process mining manifesto. In: Daniel F, Barkaoui K, Dustdar S, eds. Proc. of the Lecture Notes in Business Information Processing. Berlin, Heidelberg: Springer-Verlag, 2011. 169−194. [doi: 10.1007/978-3-642-28108-2_19]

[19] Van der Aalst WMP. Extracting event data from databases to unleash process mining. In: Brocke JV, Schmiedel T, eds. Proc. of the BPM—Driving Innovation in a Digital World. Springer Int'l Publishing, 2015. 39−47. [doi: 10.1007/978-3-319-14430-6_8]

[20] Van der Aalst WMP, Weijters T, Maruster L. Workflow mining: Discovering process models from event logs. IEEE Trans. on Knowledge and Data Engineering, 2004,16(9):1128−1142. [doi: 10.1109/TKDE.2004.47]

[21] Weijters AJMM, van der Aalst WMP, De Medeiros AKA. Process mining with the heuristics miner-algorithm. BETA Working Paper Series 166, Eindhoven: BETA Publisher in Eindhoven University of Technology, 2006.

[22] Dongen BFV, Busi N, Pinna G, van der Aalst WMP. An iterative algorithm for applying the theory of regions in process mining. In: Reisig W, Hee KV, Siedlce KW, eds. Proc. of the Workshop on Formal Approaches to Business Processes and Web Services. Publishing House of University of Podlasie, 2007. 36−55.

[23] Bergenthum R, Desel J, Lorenz R, *et al*. Process mining based on regions of languages. In: Alonso A, Dadam P, Rosemann M, eds. Proc. of the Int'l Conf. on Business Process Management (BPM 2007). Berlin, Heidelberg: Springer-Verlag, 2007. 375−383. [doi: 10.1007/978-3-540-75183-0_27]

[24] De Medeiros AKA, Weijters AJMM, van der Aalst WMP. Genetic process mining: An experimental evaluation. Data Mining and Knowledge Discovery, 2007,14(2):245−304. [doi: 10.1007/s10618-006-0061-7]

[25] Günther CW, van der Aalst WMP. Fuzzy mining—Adaptive process simplification based on multi-perspective metrics. In: Alonso A, Dadam P, Rosemann M, eds. Proc. of the Int'l Conf. on Business Process Management (BPM 2007). Berlin, Heidelberg: Springer-Verlag, 2007. 328−343. [doi: 10.1007/978-3-540-75183-0_24]

[26] Werf JMVD, Dongen BFV, Hurkens CAJ, Serebrenik A. Process discovery using integer linear programming. Fundamenta Informaticae, 2009,94(3-4):387−412. [doi: 10.3233/FI-2009-136]

[27] Cook JE, Wolf AL. Discovering models of software processes from event-based data. ACM Trans. on Software Engineering and Methodology (TOSEM), 1998,7(3):215−249. [doi: 10.1145/287000.287001]

[28] Rubin V, Günther CW, van der Aalst WMP, Kindler E, Dongen BFV, Schäfer W. Process mining framework for software processes. In: Wang Q, Pfahl D, Raffo DM, eds. Proc. of the Int'l Conf. on Software Process, Software Process Dynamics and Agility. Berlin, Heidelberg: Springer-Verlag, 2007. 169−181. [doi: 10.1007/978-3-540-72426-1_15]

[29] Garg PK, Bhansali S. Process programming by hindsight. In: Proc. of the 14th Int'l Conf. on Software Engineering. 1992. 280−293. [doi: 10.1145/143062.143128]

[30] Senin P. Software trajectory analysis: An empirically based method for automated software process discovery. Technical Report, 09-09, CSDL, 2009. http://csdl.ics.hawaii.edu/techreports/09-09/09-09.pdf

[31] Valle A, Portela E, Loures ER, Cestari JM. A framework for applying process mining techniques in software process assessments. In: Proc. of the IIE Annual Conf. 2014. 1339−1347.

[32] Kindler E, Rubin V, Schäfer W. Incremental workflow mining based on document versioning information. In: Li M, Boehm B, Osterweil LJ, eds. Proc. of the Unifying the Software Process Spectrum (SPW 2005). LNCS 3840, Berlin, Heidelberg: Springer-Verlag, 2006. 287−301. [doi: 10.1007/11608035_25]

[33] Jans M, Werf JMVD, Lybaert N, Vanhoof K. A business process mining application for internal transaction fraud mitigation. Expert Systems with Applications, 2011,38(10):13351−13359. [doi: 10.1016/j.eswa.2011.04.159]

[34] Castillo RP, Weber B, Pinggera J, Zugal S, Guzman GRD, Piattini M. Generating event logs from non-process-aware systems enabling business process mining. Enterprise Information Systems, 2011,5(3):301−335. [doi: 10.1080/17517575.2011.587545]

[35] Poggi N, Muthusamy V, Carrera D, Khalaf R. Business process mining from E-commerce Web logs. In: Daniel F, Wang J, Weber B, eds. Proc. of the 11th Int'l Conf. on Business Process Management. LNCS 8094, Berlin, Heidelberg: Springer-Verlag, 2013. 65−80. [doi: 10.1007/978-3-642-40176-3_7]

[36] Leemans M, van der Aalst WMP. Process mining in software systems: Discovering real-life business transactions and process models from distributed systems. In: Proc. of the 18th Int'l Conf. on Model Driven Engineering Languages and Systems (MODELS). ACM/IEEE, 2015. 44−53. [doi: 10.1109/MODELS.2015.7338234]

[37] Lohmann N, Verbeek E, Dijkman R. Petri net transformations for business processes—A survey. Trans. on Petri Nets and Other Models of Concurrency II, 2009,5460:46−63. [doi: 10.1007/978-3-642-00899-3]

[38] Cortadella J, Kishinevsky M, Lavagno L, Yakovlev A. Deriving Petri nets from finite transition systems. IEEE Trans. on Computers, 1998,47(8):859−882. [doi: 10.1109/12.707587]

[39] Lopezgrao JP, Merseguer J, Campos J. From UML activity diagrams to stochastic Petri nets: Application to software performance engineering. Workshop on Software and Performance, 2004,29(1):25−36. [doi: 10.1145/974044.974048]

[40] Reisig W. Petri Nets: An Introduction. Berlin, Heidelberg: Springer-Verlag, 1985. [doi: 10.1007/978-3-642-69968-9]

[41] Yuan CY. Petri Net Applications. Beijing: Science Press, 2013 (in Chinese).

[42] 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).

[43] Yang JW, Chen XO. A semi-structured document model for text mining. Journal of Computer Science and Technology, 2002,17(5): 603−610. [doi: 10.1007/BF02948828]

[44] Yang JW, Cheung WK, Chen XO. Learning element similarity matrix for semi-structured document analysis. Knowledge and Information Systems, 2009,19(1):53−78. [doi: 10.1007/s10115-008-0138-2]

[45] Yoon JP, Raghavan V, Chakilam V. BitCube: A three-dimensional bitmap indexing for XML documents. In: Proc. of the 13th Int'l Conf. on 2001 Scientific and Statistical Database Management. 2001. 158−167. [doi: 10.1023/A:1012861931139]

[46] Nayak R, Xu S. XCLS: A fast and effective clustering algorithm for heterogenous XML documents. In: Proc. of the Knowledge Discovery and Data Mining. 2006. 292−302. [doi: 10.1007/11731139_35]

[47] Tran T, Nayak R, Bruza P. Combining structure and content similarities for XML document clustering. In: Proc. of the 7th Australasian Data Mining Conf., Vol.87. ACM Press, 2008. 219−226.

[48] Algergawy AAA. Management of XML data by means of schema matching [Ph.D. Thesis]. Otto von Guericke University Magdeburg, 2010.

[49] Bezdek JC. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Springer US, 1981. [doi: 10.1007/978-1-4757-0450-1]

[50] Pal NR, Pal SK. Entropy: A new definition and its applications. IEEE Trans. on Systems, Man and Cybernetics, 1991,21(5): 1260−1270. [doi: 10.1109/21.120079]

[51] Hung WL, Yang MS. Similarity measures of intuitionistic fuzzy sets based on Hausdorff distance. Pattern Recognition Letters, 2004,25(14):1603−1611. [doi: 10.1016/j.patrec.2004.06.006]

[52] Singh VP, Asce F. Hydrologic synthesis using entropy theory: Review. Journal of Hydrologic Engineering, 2011,16(5):421−433. [doi: 10.1061 /(ASCE)HE.1943-5584.0000332]

[53] De Medeiros AKA, Dongen BFV, Weijters AJMM. Process mining: Extending the $\alpha$-algorithm to mine short loops. BETA Working Paper Series, WP 113, Eindhoven: Eindhoven University of Technology, 2004.

[54] Leemans SJJ, Fahland D, van der Aalst WMP. Discovering block-structured process models from incomplete event logs. In: Ciardo G, Kindler E, eds. Proc. of the Int'l Conf. on Applications and Theory of Petri Nets and Concurrency. LNCS, Cham: Springer-Verlag, 2014. 311−329. [doi: 10.1007/978-3-319-07734-5_6]

[55] Alves WMPVD, De Medeiros AKA, Weijters AJMM. Genetic process mining. In: Ciardo G, Darondeau P, eds. Proc. of the Applications and Theory of Petri Nets. LNCS 3536, Berlin: Springer-Verlag, 2005. 48−69. [doi: 10.1007/11494744_5]

[56] Zaki MJ. SPADE: An efficient algorithm for mining frequent sequences. Machine Learning, 2001,42(1-2):31−60. [doi: 10.1023/A:1007652502315]

[57] Zimmermann A. Understanding episode mining techniques: Benchmarking on diverse, realistic, artificial data. Intelligent Data Analysis, 2014,18(5):761−791. [doi: 10.3233/IDA-140668]

[58] Peterson JL. Petri Net Theory and the Modeling of Systems. Prentice-Hall, 1981.

**附中文参考文献**:

[12] 郁抒思,周水庚,关佶红.软件工程数据挖掘研究进展.计算机科学与探索,2012,6(1):1−31. [doi: 10.3778/j.issn.1673-9418.2012.01.001]

[16] 何克清,李兵,马于涛,黄贻望.大数据时代的软件工程关键技术.中国计算机学会通讯,2014,10(3):8−18.

[17] 王千祥,梅宏.大数据与软件工程.中国计算机学会通讯,2014,10(3):6−7.

[41] 袁崇义.Petri 网应用.北京:科学出版社,2013.

[42] 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. [doi: 10.13196/j.cims.2016.02.006]

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

**何臻力**(1987－),男,博士,讲师,主要研究领域为云计算,大数据.

**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过程,形式化方法.

**于倩**(1975－),女,博士,讲师,CCF 专业会员,主要研究领域为软件工程.

**莫启**(1986－),男,博士,讲师,主要研究领域为业务过程管理.

**王一荃**(1983－),女,博士生,助理研究员,主要研究领域为软件过程,Petri 网.