Table 6 Test data for HAIA experiments on many to many relation between cases

表 6 HAIA 多对多关系实验数据

测试类	匹配关系	测试组	实例数	事件数	事件属性数
T1	0型、I型	G4	869 vs. 1 332	4 083 vs. 10 169	7
		G5	2 702 vs. 4 127	12 764 vs. 32 161	
		G6	5 365 vs. 8 238	24 702 vs. 61 713	
T2	0型、II型	G7	1 115 vs. 595	4 519 vs. 6 528	7
		G8	2 218 vs. 1 181	12 963 vs. 9 167	
		G9	4 479 vs. 2 349	26 492 vs. 17 890	
T3	0型、I型、II型、III型	G10	1 243 vs. 1 108	7 530 vs. 6 712	7
		G11	3 626 vs. 3 418	18 749 vs. 17 783	
		G12	7 061 vs. 6 813	43 169 vs. 38 711	

表 6 中,测试数据分为 3 大类:T1,T2,T3,分别包含了 I 型关系、II 型和 III 型关系的事件日志,每类数据包含 3 组数据,均包含少量随机产生的噪声.3 组数据的规模基本上成倍数增长.实验中,初始种群分别使用随机方法和启发式方法生成,规模均为 100,免疫进化的停止条件设置为连续 50 代种群中个体亲和度最大值不再提高.

采用随机方法生成初始种群的实验,主要目的是验证启发式方法对免疫进化的影响.按照包含 I 型、II 型和 III 型关系的各匹配矩阵的特点,个体中每个匹配关系均采用完全随机方式来确定.采用随机方法生成初始种群的实验结果见表 7.采用启发式方法生成初始种群的实验结果见表 8.

Table 7 Test results of HAIA experiments on many to many relation between cases (randomly generated initial population)

表 7 HAIA 多对多关系实验结果(随机生成初始种群)

测试类别	测试组	种群初始化平均执行时间	种群进化平均执行时间(s)	融合成功率(%)
T1	G4	4.8	81.5	91.98
	G5	15.5	595.6	91.74
	G6	32.0	2 223.6	91.56
T2	G7	3.9	60.6	92.04
	G8	9.7	194.6	91.88
	G9	18.0	512.4	91.73
T3	G10	5.3	116.3	91.25
	G11	18.8	625.3	91.11
	G12	35.7	2 414.3	91.07

Table 8 Test results of HAIA experiments on many to many relation between cases (heuristic to generate initial population)

表 8 HAIA 多对多关系实验结果(启发式生成初始种群)

测试类别	测试组	种群初始化平均执行时间	种群进化平均执行时间(s)	融合成功率(%)
T1	G4	68.8	8.5	94.08
	G5	400.6	73.4	93.91
	G6	1 603.6	390.0	93.80
T2	G7	22.8	4.6	94.12
	G8	92.5	22.6	93.81
	G9	346.5	61.4	93.78
T3	G10	77.3	16.1	93.88
	G11	415.8	152.3	93.91
	G12	1 654.1	491.5	93.67

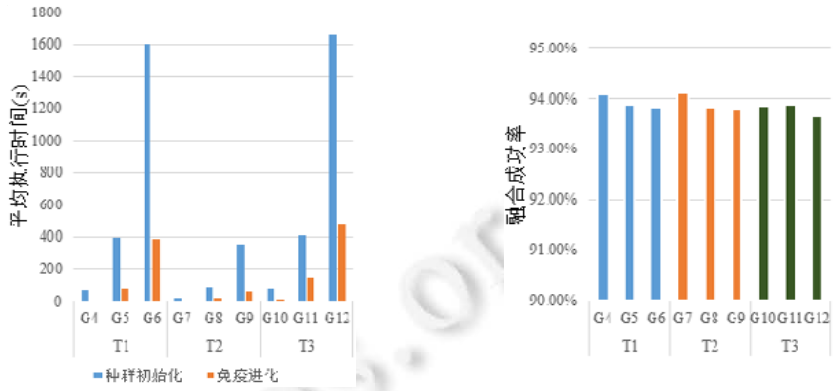
从表 7 和表 8 的实验结果对比来看,采用随机方法生成初始种群平均执行时间远小于采用启发式方法生成初始种群.这是由于实例间相同属性值比率 $ratio_{attr}(\omega_x,\omega_y)$的计算是要对$\omega_x$ 和ω_y 的所有事件的所有属性进行比较,因而 $ratio_{attr}(\omega_x,\omega_y)$的计算是非常耗时的.

然而,随机方法生成初始种群的免疫进化消耗的平均执行时间远大于采用启发式方法生成初始种群的进化平均执行时间.综合来看,采用启发式方法生成初始种群的免疫算法总耗时要小于采用随机方法生成初始种群免疫算法.这说明采用启发式方法产生初始种群,能够有效地提高免疫进化的效率,加快进化收敛.同时,从表 7 和表 8 也可以看到,采用启发式方法生成初始种群的融合成功率比采用随机方法生成初始种群的融合成功率

略高.

从表 8 中可知:在日志中包含 I 型关系、II 型和 III 型关系的 3 类测试中,HAIA 融合成功率都在 93%以上.在整个算法中,种群初始化的时间开销要远远高于进化阶段的时间开销.这个实验结果和第 2.5 节算法复杂度的分析结论一致,启发式方法生成的随机种群的复杂度比种群进化计算的复杂度要高.

同样需要注意的是:无论事件日志中包含 I 型关系、II 型还是 III 型匹配关系,随着事件日志规模的不断扩大,HAIA 的种群初始化和免疫进化的效率都呈恶化趋势;HAIA 融合成功率相对稳定,但仍呈现逐渐下降趋势,如图 7 所示.



(a) HAIA 多对多关系融合效率 (b) HAIA 多对多关系融合成功率

Fig.7 Trends of merging quality and efficiency of HAIA

图 7 HAIA 融合质量和融合效率趋势

相对于 0 型匹配关系的融合问题,I 型、II 型和 III 型匹配关系的实例间关系比较复杂.因此,当事件日志规模增大时,日志中包含的实例间关系复杂性均增加,导致融合效率呈下降趋势.图 6(a)中,在日志规模增长大体相当的情况下,包含多对多关系日志的融合效率恶化更迅速,这也表明日志规模增大,多对多关系融合问题的复杂性比其他类型的匹配关系更加严重.

上述实验表明:(1) HAIA 能够较好地实现包含多对多匹配关系的日志融合;(2) 在保证融合成率的情况下,HAIA 算法比 AIA 算法能够更快地收敛(针对一对一匹配关系);(3) 启发式方法生成初始种群,能够提高 HAIA 的免疫进化的效率;(4) 随着日志规模的增大,日志中匹配关系的复杂度升高,HAIA 融合性能趋于恶化.

4 讨 论

第 2.5 节的 HAIA 算法复杂度分析说明,HAIA 的性能随着事件日志的规模增大而趋于下降.第 3 节的实验结果进一步表明:两个待融合的日志中包含的匹配关系类型越复杂,HAIA 的性能下降得越快.如何面对大规模日志数据,有效地提高日志融合效率,是日志融合技术得到实际应用需要解决的重要问题.

如今,诸如多核计算、集群计算、网格计算、云计算等分布式计算系统被广泛应用于提高计算性能和可扩展性,分布式聚类[19,20]、分布式关联规则挖掘[21]等分布式数据挖掘技术也被用于提高数据挖掘性能.分布式流程挖掘技术的研究也见于文献[3,22,23].同样地,日志数据融合也可以采用分布式技术来提高融合性能.本节将对分布式流程日志数据融合技术进行探讨.在此,本文不关注分布式融合的实现细节,而是重点讨论分布式 HAIA 中的数据划分问题.

4.1 种群初始化

HAIA 的启发式方法构建初始种群的个体的核心是计算日志间两两实例的相同属性值比率 $ratio_{attr}(\omega_x,\omega_y)$.从算法分析和实验结果看,$ratio_{attr}(\omega_x,\omega_y)$的计算是非常耗时的.从 $ratio_{attr}(\omega_x,\omega_y)$的计算方式看,$ratio_{attr}(\omega_x,\omega_y)$的

计算仅与ω_x和ω_y两个实例中的事件相关,与其他实例无关.因而可以考虑将日志划分为多个不同的实例子集,每个子集分配给不同的计算节点进行 $ratio_{attr}(\omega_x,\omega_y)$的计算,通过这种分布式计算方式提高初始化的效率.

日志中的每个实例需要与另一个日志的所有实例进行 $ratio_{attr}$ 的计算.假设有 n 个计算节点,一种方案是考虑将规模较大的日志划分为 n 个不同子集,分别分配到 n 个计算节点上,而将规模较小的日志(实例数量较少的日志)的所有实例在每个节点上复制,计算的结果可以按照矩阵的行或列的形式进行融合.日志划分为子集时,可以将规模(实例中包含事件数量)相当的实例均匀划分到各子集,使各子集的 $ratio_{attr}$ 计算量尽可能地均衡.如图 8 所示,日志 logA 根据各实例的规模被划分为两个实例子集{case1,case4,case5},{case2,case3,case6}.这两个子集分别分配给两个计算节点,同时,将logB所有实例复制到这两个计算节点,分别在这两个节点上并行计算得到两个子匹配矩阵,最后,将这两个子矩阵合并,得到初始种群中的一个匹配矩阵.这种方案在子匹配矩阵计算过程中,节点间不需要进行数据交换,因而分布式计算中的通信开销小.但是对于"大数据"级的日志,将大规模日志的所有实例复制到每个节点并不适合.

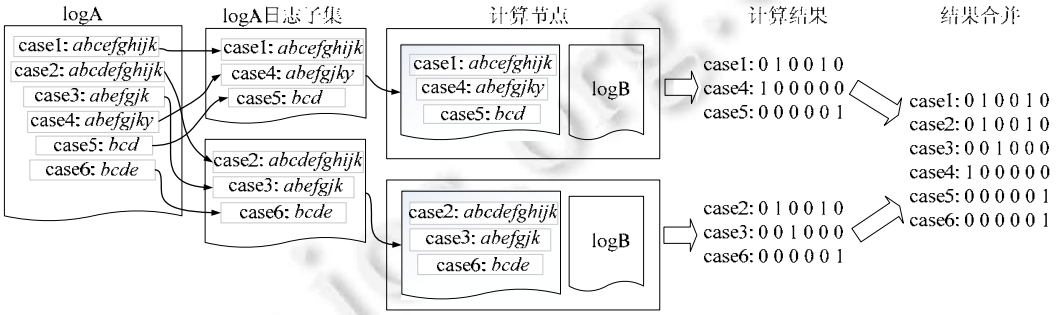


Fig.8 Partitioning the log according to the cases size

图 8 根据实例规模的日志划分

针对"大数据"级的日志,可以将两个待融合日志均按照实例规模划分为n个子实例集合,分别分配给n个计算节点.在分布式计算过程中,规模较小的日志的子实例集合在各个节点间进行交换,直至两个日志所有实例间均进行了计算.如图 9 所示,日志 logA 根据各实例的规模被划分为两个实例子集{case1,case4,case5},{case2,case3,case6},logB 则划分为两个实例子集{case1′,case2′,case4′}和{case3′,case5′,case6′}.{case1,case4,case5}与{case1′,case2′,case4′}完成匹配关系计算后,节点2将{case3′,case5′,case6′}数据交换到节点1,{case1,case4,case5}再与{case3′,case5′,case6′}进行匹配关系计算.

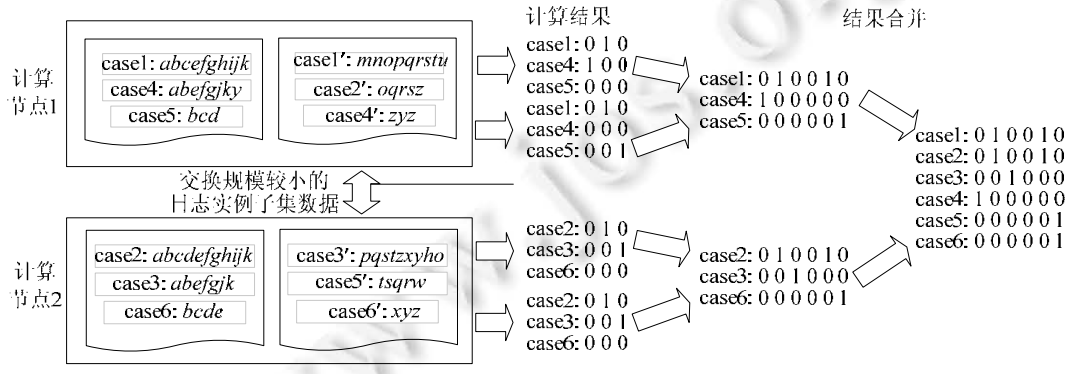


Fig.9 Partitioning the log as "big data"

图 9 针对"大数据"级日志的划分

同样地,完成{case2,case3,case6}与{case3′,case5′,case6′},{case1′,case2′,case4′}的计算.计算结果分两步进行合并,logA 的实例子集{case1,case4,case5}与 logB 的{case1′,case2′,case4′},{case3′,case5′,case6′}计算得到两个匹配关系子矩阵,这两个匹配关系子矩阵首先在节点 1 合并.同样地,计算{case2,case3,case6}与 logB 在节点 2 上合并后的匹配关系子矩阵.最后,将节点 1 和节点 2 上的匹配关系子矩阵合并,得到初始种群中的一个匹配矩阵.这种方案节点间需要进行数据交换,网络通信开销大.

4.2 分布式免疫进化

HAIA 免疫进化的核心是每一代种群中个体亲和度的计算和排序.个体亲和度的计算与种群中其他个体没有关系,因此,种群中个体亲和度计算适于采用分布式方式实现并行.可以将种群划分为 n 个子种群,分别在 n 个节点独立进行免疫进化,节点间通过各子种群免疫进化结果再进行亲和度比较,选出亲和度最高的个体作为最终解.需要进一步考虑的问题是:1) 如何划分子种群;2) 子种群间如何交换其“最佳”个体.

- 种群划分

子种群的划分需要考虑的一个重要因素是如何避免子种群中个体的亲和度均偏“高”或偏“低”,导致子种群在进化中容易陷入早熟.因此,需要亲和度相近的个体均衡分布在各子种群中.在两个匹配矩阵间,相同的匹配关系数量越多,表明这两个匹配矩阵的亲和度值越接近,这里称作“亲近度”.假设计算节点为 n,根据个体间亲近度将种群中的个体划分为若干个聚类,再将各个聚类中的个体均匀分配为 n 个子种群.这样,可以保证子种群的个体亲和度是均衡的.

匹配矩阵是 0-1 二值矩阵,可以考虑两个矩阵对应项之间进行异或计算,即 $a_{ij}\otimes b_{ij}$,异或计算得到的矩阵中值为 1 的项越少,表明这两个矩阵的相同的匹配关系数量越多,它们的近亲度越高近.如图 10 所示,匹配矩阵 A 分别与 B,C 进行对应项异或计算,得到矩阵 D 和 E.其中,D 中为 1 的项数量为 1,E 中为 4,则认为个体 A 和个体 B 的亲近度高于 A 和 C 的亲近度.可以发现,异或计算得到的矩阵所有项值的和,即 $\sum_M a_{ij}$, 就是该矩阵中为 1 的项的数量,因此,可以通过比较 D 和 E 的所有项值的和来判定 A 与 B 的亲近度和 A 和 C 的亲近度的高低.

$$M\otimes N \qquad B=\begin{bmatrix}0&1&1\\0&0&0\\1&0&0\end{bmatrix} \qquad C=\begin{bmatrix}1&0&1\\0&1&0\\1&0&0\end{bmatrix}$$

$$A=\begin{bmatrix}1&0&1\\0&1&0\\1&0&1\end{bmatrix} \qquad D=\begin{bmatrix}0&0&0\\0&0&0\\0&0&1\end{bmatrix} \qquad E=\begin{bmatrix}1&1&0\\0&1&0\\0&0&1\end{bmatrix}$$

$\sum_D a_{ij}=1<\sum_E a_{ij}=4$,因此,$A$的亲和度与$B$更接近

Fig.10 XOR between match matrixces

图 10 匹配矩阵间的异或计算

- 个体迁移

子种群间的“最佳”个体的交换方式,即分布式 HAIA 的个体迁移,需要考虑迁移个体的类型、迁移个体的规模、接受迁移个体方式、迁移间隔.

分布式 HAIA 中,子种群间迁移的个体既可以是各子种群亲和度排名高的个体,也可以考虑选取随机的个体做迁移.免疫系统的克隆选择理论的基本思想是,只有那些能够识别抗原的细胞才能增殖.因此,选择亲和度“最佳”的个体进行迁移是较好的策略.

在分布式进化算法中,子种群的个体迁移往往采用子种群间的广播方式,即按照设定的迁移规模向所有其他子种群交换个体.而在 HAIA 中设置了免疫记忆库机制,用于存储的是免疫进化过程中亲和度排在前列的个体,这种保证多样性的目的与个体迁移的目的相同.因而在分布式 HAIA 中,除了子种群间的广播方式以外,可以考虑采用以记忆库为中介的个体迁移方式,所有子种群将本种群中排名前 n 位的个体提交给记忆库,按照记忆

库的“更新”策略,子种群间进行迁移个体“竞争”,最后,将亲和度在全局排名前列的个体存储在记忆库中.各子种群则按照记忆库“复制”策略从记忆库获取“最佳”个体来接受迁移的个体,用于产生下一代子种群.关于迁移间隔,在 HAIA 的进化中,记忆库的更新及复制在每一代都发生,考虑个体的迁移在子种群每一代都进行.有研究表明:在分布式进化计算中,小的迁移间隔可能发生某些子种群主宰其他子种群的情况而导致全局多样性的降低,大的迁移间隔会降低进化收敛的速度[24].在分布式 HAIA 中,记忆库最终存储的是各子群通过“竞争”而保留下来的亲和度排名“最佳”的个体集合,加上初始种群均衡划分策略和模拟退火机制的采用,因而可以极大地降低出现某些子种群主宰其他子种群的机会.

5 结 论

实际业务中的流程灵活性、融合所需信息的缺失以及日志本身的“噪声”,给流程挖掘日志融合带来了挑战.本文对事件日志融合问题进行了形式化定义,指出该问题是一个搜索优化问题,并提出了一种基于混合人工免疫算法的日志融合方法 HAIA.这种方法以人工免疫系统的克隆选择理论为基础,通过免疫进化获得“最佳”解.在免疫进化的每一代,使用两个实例级别的因素,流程执行路径出现的频次和流程实例间的时间匹配关系,分别从“量”匹配和“时间”匹配两个维度对进化过程中的个体(匹配矩阵)进行评价,通过克隆、变异操作选择保留亲和度高的个体,直至获得“最佳”个体.实验结果表明:(1) HAIA 支持包含复杂流程实例间匹配关系的日志融合;(2) 启发式方法生成初始种群,能够加快免疫进化的搜索性能;(3) 免疫记忆库和模拟退火机制的引入能够保持种群的多样性,减少陷入早熟陷进的机会.

针对大规模流程日志的融合性能趋于恶化的问题,本文还讨论了分布式日志融合中的数据划分问题:针对种群初始化,提出了以实例规模进行数据划分的方法;针对免疫进化,提出了以匹配矩阵间的亲近度为基础的聚类方法的子种群划分策略以及以免疫记忆库为媒介的子种群间个体迁移方法.但是对不断增大的日志规模来说,目前“离线”方式的日志融合方法的性能会受到日志数据的存储方案的影响,而且在时效性方面比较差.流式的日志融合方法是未来进一步研究的方向之一.

References:

[1] Beheshti SMR, Benatallah B, Sakr S, Grigori D, Motahari-Nezhad HR, Barukh MC, Gater A, Ryu SH. Process Analytics: Concepts and Techniques for Querying and Analyzing Process Data. Springer Int'l Publishing, 2016. [doi: 10.1007/978-3-319-25037-3]

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

[3] van der Aalst WMP. Process Mining: Data Science in Action. Berlin, Heidelberg: Springer-Verlag, 2016. [doi: 10.1007/978-3-662-49851-4]

[4] Zikopoulos P, Eaton C. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media, 2011.

[5] Weijters AJMM, Ribeiro JTS. Flexible heuristics miner (FHM). In: Proc. of the Computational Intelligence and Data Mining. IEEE, 2011. 310–317. [doi: 10.1109/CIDM.2011.5949453]

[6] Medeiros AK, Weijters AJ, 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]

[7] van der Aalst WMP. Distributed process discovery and conformance checking. In: Proc. of the Int'l Conf. on Fundamental Approaches to Software Engineering. Springer-Verlag, 2012. 1–25. [doi: 10.1007/978-3-642-28872-2_1]

[8] van der Aalst WMP, Adriansyah A, Van Dongen B. Replaying history on process models for conformance checking and performance analysis. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, 2012,2(2):182–192. [doi: 10.1002/widm.1045]

[9] Claes J, Poels G. Merging computer log files for process mining: An artificial immune system technique. In: Proc. of the Business Process Management Workshops. Berlin, Heidelberg: Springer-Verlag, 2011. 99–110. [doi: 10.1007/978-3-642-28108-2_9]

[10] Pérez-Castillo R, Weber B, de Guzmán IG, Piattini M, Pinggera J. Assessing event correlation in non-process-aware information systems. Software & Systems Modeling, 2014,13(3):1117–1139. [doi: 10.1007/s10270-012-0285-5]

[11] Motahari-Nezhad HR, Saint-Paul R, Casati F, Benatallah B. Event correlation for process discovery from Web service interaction logs. The VLDB Journal, 2011,20(3):417−444. [doi: 10.1007/s00778-010-0203-9]

[12] Claes J, Poels G. Integrating computer log files for process mining: A genetic algorithm inspired technique. In: Proc. of the Advanced Information Systems Engineering Workshops. Berlin, Heidelberg: Springer-Verlag, 2011. 282−293. [doi: 10.1007/978-3-642-22056-2_30]

[13] Claes J, Poels G. Merging event logs for process mining: A rule based merging method and rule suggestion algorithm. Expert Systems with Applications, 2014,41(16):7291−7306. [doi: 10.1016/j.eswa.2014.06.012]

[14] Murata T. Petri nets: Properties, analysis and applications. Proc. of the IEEE, 1989,77(4):541−580. [doi: 10.1109/5.24143]

[15] Eiben AE, Smith JE. Introduction to Evolutionary Computing. Heidelberg: Springer-Verlag, 2003. [doi: 10.1007/978-3-662-44874-8]

[16] Burke EK, Kendall G. Search Methodologies-Introductory Tutorials in Optimization and Decision Support Techniques. 2nd ed., New York: Springer-Verlag, 2014. [doi: 10.1007/978-1-4614-6940-7]

[17] Shu WN. Artificial immune algorithm optimization and its key problems research [Ph.D. Thesis]. Wuhan: Wuhan University, 2013 (in Chinese with English abstract).

[18] Fu WY, Ling CD. Brownian motion based simulated annealing algorithm. Journal of Computer, 2014,6(37):1301−1308 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2014.01301]

[19] Visalakshi NK, Thangavel K. Distributed data clustering: A comparative analysis. In: Proc. of the Foundations of Computational, Intelligence, Vol.6. Berlin, Heidelberg: Springer-Verlag, 2009. 371−397. [doi: 10.1007/978-3-642-01091-0_16]

[20] Singh D, Gosain A. A comparative analysis of distributed clustering algorithms: A survey. In: Proc. of the Int'l Symp. on Computational and Business Intelligence. IEEE, 2013. 165−169. [doi: 10.1109/ISCBI.2013.40]

[21] Sawant V, Shah K. A survey of distributed association rule mining algorithms. Journal of Emerging Trends in Computing and Information Sciences, 2014,5(5):391−398.

[22] Alhajj R, Rokne J. Distributed Process Mining, Encyclopedia of Social Network Analysis and Mining. New York: Springer-Verlag, 2014. 400−403. [doi: 10.1007/978-1-4614-6170-8_100682]

[23] Bratosin CC. Grid architecture for distributed process mining. Technische Universiteitndhoven, 2011. [doi: 10.6100/IR699500]

[24] Skolicki Z, De Jong K. The influence of migration sizes and intervals on island models. In: Proc. of the 7th Annual Conf. on Genetic and Evolutionary Computation. ACM Press, 2005. 1295−1302. [doi: 10.1145/1068009.1068219]

附中文参考文献:

[17] 舒万能.人工免疫算法的优化和关键问题研究[博士学位论文].武汉:武汉大学,2013.

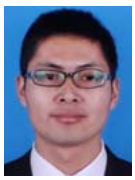
[18] 傅文渊,凌朝东.布朗运动模拟退火算法.计算机学报,2014,6(37):1301−1308. [doi: 10.3724/SP.J.1016.2014.01301]

徐杨(1970－),男,湖北武汉人,博士,讲师,主要研究领域为分布式计算,流程建模,流程分析.

汤德佑(1976－),男,博士,副教授,CCF 专业会员,主要研究领域为数据起源,数据库,高性能计算.

袁峰(1977－),男,博士,副研究员,主要研究领域为物联网,云计算,大数据.

李东(1970－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为大数据与云计算,业务流程管理.



林琪(1991－),男,硕士,主要研究领域为流程建模,流程分析.