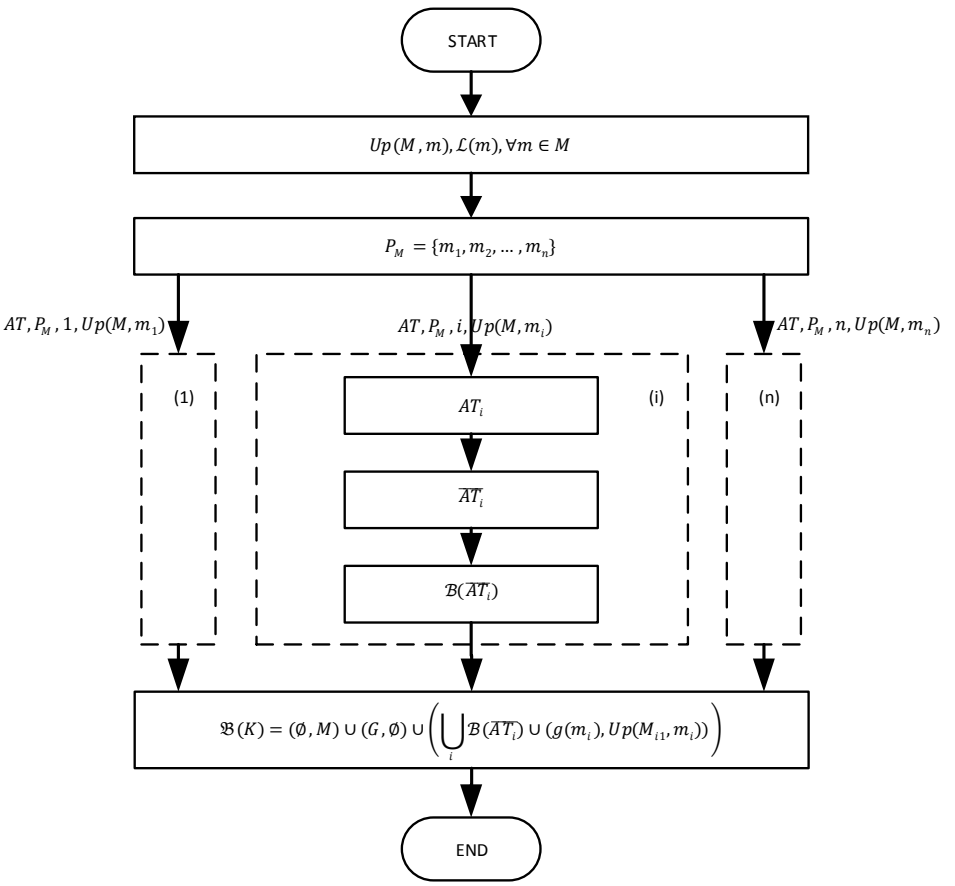


Fig.5 The flowchart of the Algorithm 4

图 5 算法 4 的流程图

对于并行的概念计算算法来说,除了考虑形式背景的拆分和概念的合并之外,使得各部分的负载得到均衡也是一个良好的并行计算算法应该具有的特性.一般来说,在将所有概念计算完成之前,形式概念的分布情况是不能预知的,因此,很难将复杂度平均的分发给各个线程[10].而在 BDAT 分解过程中,属性拓扑的分解和子属性拓扑的约简算法限制了最终拓扑的大小.对于以伴生属性为中心的约简子属性拓扑中,其中的属性顶点数不会超过#*SupAttr*;对于以顶层属性为中心的约简子属性拓扑中,其中的属性顶点数必小于#*SupAttr*,并且以有序属性集合的顺序,其约简子属性拓扑的大小是越来越小的.考虑到形式概念数大致与形式背景的大小成指数型关系增长[29],因此,我们可以大致地认为,约简子属性拓扑的概念计算复杂度随着顶层属性的排序逐渐递减.则在线程有限的情况下,按照有序属性集合中的顺序,首先计算以伴生属性为中心的子属性拓扑,然后通过适当的分配顶层属性为中心的子属性拓扑,可以使不同线程的负载尽可能均衡,这使得 BDAT 算法更适合并行计算.

本节中,使用生物和水的属性拓扑作为数据源.在第 3.2 节中,已经求解出以每个属性为中心的各个约简子属性拓扑(如图 4 所示),只需计算出每个约简子属性拓扑的所有概念,再根据定理 3,由 8 个中心属性构成的 8 个概念,通过简单的并集运算,即可得到没有重复概念和伪概念的全体形式概念集合(见表 4).表 4 中列出了生物和水的形式背景下使用 BDAT 得到的全体概念,所得概念是正确的且没有遗漏和丢失,其中,第 12 个概念的外延为{6,7,8}={芦苇,豆,玉米},内涵为{*c*,*d*}={在陆地上生活,有叶绿素},在当前的形式背景中,则可以将该概念理解为陆生植物.同样的,在该背景下的两栖生物对应于第 16 个概念,其外延为{3,6}={青蛙,芦苇},内涵为{*b*,*c*}={在水中生活,在陆地上生活}.

Table 4 All the concepts of Living Beings and Water computed by BDAT algorithm

表 4 使用 BDAT 算法计算出的生物和水的形式概念

编号	外延	内涵	来源	编号	外延	内涵	来源
1	4	*c*,*g*,*h*,*i*	中心属性 *i*	11	5,6,7,8	*d*	中心属性 *d*
2	7	*c*,*d*,*e*	中心属性 *e*	12	6,7,8	*c*,*d*	顶层属性 $\overline{AT_5}$
3	5,6,8	*d*,*f*	中心属性 *f*	13	1,2,3,4	*g*	中心属性 *g*
4	6,8	*c*,*d*,*f*	顶层属性 $\overline{AT_3}$	14	1,2,3	*b*,*g*	顶层属性 $\overline{AT_6}$
5	5,6	*b*,*d*,*f*	顶层属性 $\overline{AT_3}$	15	1,2,3,5,6	*b*	中心属性 *b*
6	6	*b*,*c*,*d*,*f*	生成概念 $\overline{AT_3}$	16	3,6	*b*,*c*	顶层属性 $\overline{AT_7}$
7	2,3,4	*g*,*h*	中心属性 *h*	17	3,4,6,7,8	*c*	中心属性 *c*
8	3,4	*c*,*g*,*h*	顶层属性 $\overline{AT_4}$	18	∅	*b*,*c*,*d*,*e*,*f*,*g*,*h*,*i*	全局概念
9	2,3	*b*,*g*,*h*	顶层属性 $\overline{AT_4}$	19	1,2,3,4,5,6,7,8	∅	全局概念
10	3	*b*,*c*,*g*,*h*	生成概念 $\overline{AT_4}$				

5 实验结果与分析

为了验证本文提出的并行形式概念计算算法的正确性,并评估并行概念计算算法的效率,实验中选取了 5 个数据集,除了典型的形式背景生物和水(living beings and water)之外,还选取了 4 组来自 UCI 的数据集: Balance scale[30],Tic tac toc[31],Mushroom[32]和 Nursery[33].这些数据集大多是多值的,因此实验前需要首先将它们转化为二值背景[1],然后经过预处理去除冗余的对象和属性,得到净化后的二值形式背景,实验中使用的各个二值形式背景的基本信息列于表 5 中.

Table 5 Information of formal contexts in discussion

表 5 实验中使用的二值形式背景的基本信息

编号	名称	对象数	属性数	复杂度	概念数
1	Living beings and water	8	8	0.41	19
2	Balance scale	625	23	0.22	2 106
3	Tic tac toc	958	28	0.34	52 717
4	Mushroom	2 744	79	0.20	47 458
5	Nursery	12 960	27	0.30	115 201

本节实验在相同的硬件和软件环境下(见表 6),测试 3 种不同的形式概念计算算法,这 3 种不同的算法是:

(1) Krajca Petr 提出的并行递归算法,PcbO;

(2) 本文提出的基于 BDAT 的并行计算算法,工作在单线程递归模式,本实验中简写为 BDAT/s;

(3) 本文提出的基于 BDAT 的并行计算算法,工作在多线程递归模式,本实验中简写为 BDAT/p.

Table 6 The hardware environment and software environment

表 6 实验的硬件和软件环境

名称	型号	核心参数
CPU/L2	Intel Core i3-3220	3.30GHz/512K
内存	Kingston	4GB/1600MHz
硬盘	Seagate	500G/7200RPM
操作系统	Windows 10	x64
编译平台	Visual Studio	2013

PCbO 是著名的形式概念计算算法之一,也被认为是最快速的算法之一[34],被广泛应用在很多领域[35,36].由于在直观图、属性偏序图等图方法中未见并行计算的相关论文,本文与第 1 种并行算法 PCbo 进行对比.PCbO 的源码来自 sourceforge 中提供的符合 GPL V2 标准的 C 语言代码,实验中将其作为本文算法的对比算法,在开源协议条款的允许下,在代码的入口和出口添加了必要的时间监测,以获取该算法的运行时间.算法执行中设置的参数均设置为作者建议的参数.

(1) 线程数设置为 4,因为代码文档中指出:通常情况下,线程数设置为 CPU 内核的 2 倍~3 倍;

(2) 其他的参数均保留其默认值.

为了更好地研究本文提出的算法,使用 C 语言编写设计了一个测试性的基于 BDAT 概念计算程序.对于属性拓扑中的完全连接图而言,任意属性的组合都将构成一个形式概念的内涵,无需进行递归调用.受限于 C 矩阵运算的复杂性,用于测试的 BDAT 算法程序在递归分解的过程中没有充分利用完全连接图的优势特征,而是将每一个拓扑分解成完全隔离的节点.因此,测试数据显示的是 BDAT 算法中最坏的情况.这个程序不但能帮我们验证本文提出的形式概念并行计算结果的正确性,同时也将通过代码监视程序运行的时间,即计算所有概念所消耗的时间,以评估本文提出算法的计算性能.

在进行实验时,考虑到算法的运行时间将会受到系统中其他进程的影响,为了减小实验的偶然误差,提高计算时间记录的准确性,每种算法分别对每个数据集进行了 10 次测试,最后对这 10 次记录的时间监视数据求取平均值,作为最终的算法运行时间(见表 7).通过实验结果比对,这 3 种算法输出的概念集合是一致的,验证了本文提出的基于 BDAT 的形式概念并行计算算法的正确性.

Table 7 The time costs in computing different formal context by dissimilar algorithms

表 7 输入不同形式背景时不同算法计算所有形式概念所消耗的时间

编号	形式背景	PCbO 并行耗时(ms)	BDAT 串行耗时(ms)	BDAT 并行耗时(ms)	BDAT 并串比(%)	BDAT 递归次数	速度提升
1	Living beings and water	0.0	0.1	0.5	500	3	–
2	Balance scale	6.4	5.3	4.1	77.36	4	35.94%
3	Tic tac toc	163.4	193.5	113.4	58.60	9	30.60%
4	Mushroom	229.0	260.3	143.8	55.24	10	37.21%
5	Nursery	360.3	1 275.9	599.1	46.96	8	–

在表 7 中,第 3 列~第 5 列分别记录了在 3 种不同算法下,计算形式概念消耗的时间,作为算法性能对比的基础数据;第 6 列的并串比描述了 BDAT 并行算法与串行算法时间的比值,可以清晰看出并行算法提升的速度;第 8 列(最后一列)清晰地描述了 BDAT 并行算法较 PCbO 并行算法提升的效果;第 7 列记录了使用 BDAT 算法完成概念计算过程中,递归调用的最大次数.正如前文所述,算法中没有利用完全图特性,因此本文算法理论上会小于等于实验记录中的递归次数,即,表 7 中记录的 BDAT 实验数据是算法的最坏情形.

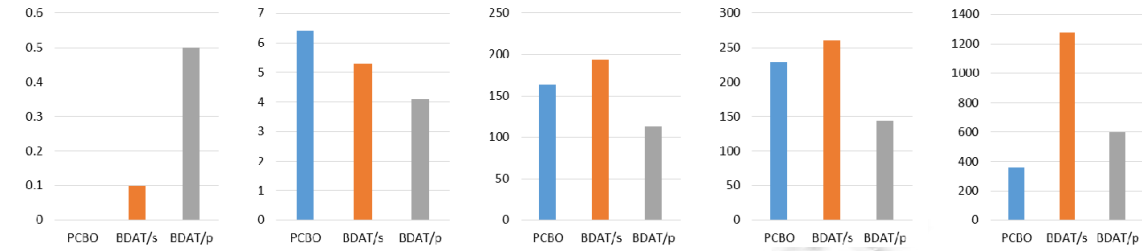
结合实验数据进行分析,可以得出以下结论.

(1) 对于较小的形式背景,如生物和水的背景,BDAT 算法慢于 PCBO,且表现出串行优于并行的现象(如图 6(a)所示).出现这种现象的根本原因是:虽然经过 BDAT 的分解和约简使子拓扑的规模得到降低(见第 3 节),但以此减少的概念计算时间并未抵消属性拓扑的构建、分解、约简和线程操作所占用的时间;

(2) 当形式背景规模较大且对象数较少时,如第 2~4 个形式背景,由于属性拓扑分解和约简算法的相互独立性,加之属性拓扑对形式背景的升维操作,本文提出的 BDAT 并行算法比 PCbO 并行算法速度提升 30%左右(如图 6(b)~图 6(d)所示).类似的,我们可以使用形式背景的转置,使本文的并行概念计算算法适应规模较大且属性数较小的形式背景;

(3) 对于 Nursery 背景,其规模较大对象数较多但只有 27 个属性,每个子拓扑的分解程度不够细小,每个线程需要处理大量的数据,因此,BDAT 算法并没有表现出优势(如图 6(e)所示).若将形式背景的转置后计算形式概念,则需要大量的线程,每个线程计算量较为细小,受到实验环境并行执行线程数的限制,没有给出并行线程数足够情形下的实验结果,因此,本文算法对此类背景的适应性有待进一步研究;

(4) 随着形式背景规模的增大,BDAT 并串比如图 7 所示,其最后将近似趋于一个常数,这个常数与 CPU 的核数有关.



(a) Living beings and water (b) Blance Scale (c) Tic tac toc (d) Mushroom (e) Nursery

Fig.6 The average time costs (ms) by different algorithms in different contexts

图 6 在不同形式背景下不同算法的计算平均时间(ms)

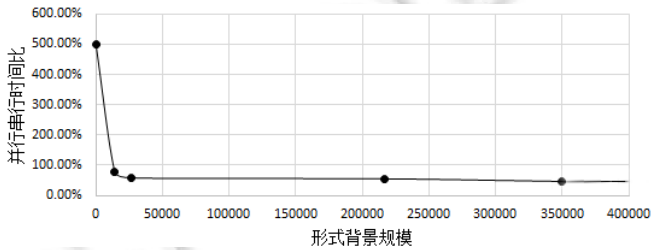


Fig.7 The chart showing the ratio of parallel time complexity and sequential with context size

图 7 并行串行时间复杂度比值随形式背景规模的曲线示意图

假定形式背景的规模相同,属性拓扑的结构将是影响计算性能的关键因素,现在已知的是顶层属性的个数将限制子属性拓扑的规模(见第 4 节).在形式背景规模达到或超过系统底层操作的影响可忽略的规模时,随着顶层属性所占比例的递增,BDAT 计算性能的曲线将呈现先上升后下降的趋势.为了讨论顶层对性能曲线的影响,设伴生属性的各个 Level 呈现随机分布,且当样本数足够多时,各 Level 层的伴生属性数据呈现均匀分布,此时,若顶层属性所占比例极小,则由于该算法是自下而上进行分解约简,虽然所得各个子拓扑规模较小,由于顶点的重复出现的概率很大,此时计算效率较低.随着顶层属性数所占比例的增大,计算效率将逐步加快.若顶层所占比率继续增大,子拓扑的平均规模也将由于顶层属性的增多而增大,此时,各个线程计算的耗时也将增大.

6 结束语

本文利用属性拓扑中对于属性间耦合表示的特性,提出了自下而上的拓扑分解方法,配合子拓扑的约简实现了各子拓扑间的完全并行运算.该方法解决了原有并行概念计算中并行算法串并交替问题,并从理论和实验中验证了本文方法的正确性,为快速概念分析提供了基础性工具.通过实验分析可知:与经典并行概念分析算法相比,本文方法具有快速、准确的特点.本文算法不但丰富了属性拓扑理论,为属性拓扑更广泛的研究提供了新的思路,使得大规模形式背景的概念分析成为了可能,同时也可应用于认知计算等相关的领域,作为基础算法为高层算法提供数据计算服务.

由于属性拓扑可以理解为广义的有向加权图,在下一步的研究中将利用属性拓扑作为媒介,将分布式图计算与形式概念计算连接起来,应用于机器学习等领域.

References:

[1] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations. Springer Science & Business Media, 2012.

[2] Bhatnagar R, Kumar L. An efficient map-reduce algorithm for computing formal concepts from binary data. In: Proc. of the 2015 IEEE Int'l Conf. on Big Data. IEEE, 2015. 1519−1528.

[3] Sun XB, Li Y, Li BX, Wen WZ. A survey of using formal concept analysis for software maintenance. Acta Electronica Sinica, 2015,43(7): 1399−1406 (in Chinese with English abstract).

[4] Shao MW, Yang HZ, Wu WZ. Knowledge reduction in formal fuzzy contexts. Knowledge-Based Systems, 2014,73:265−275.

[5] Li J, Mei C, Wang J, Zhang X. Rule-Preserved object compression in formal decision contexts using concept lattices. Knowledge-Based Systems, 2014,71:435−445.

[6] Li X, Luo J, Shi A. An improved data mining algorithm based on concept lat-tice. In: Proc. of the 2nd Int'l Conf. on Computer Science and Electronics Engineering. Atlantis Press, 2013.

[7] Kaytoue M, Codocedo V, Buzmakov A, Baixeries J, Kuznetsov SO, Napoli A. Pattern structures and concept lattices for data mining and knowledge processing. In: Proc. of the Machine Learning and Knowledge Discovery in Databases. Springer Int'l Publishing, 2015. 227−231.

[8] Singh PK, Kumar CA, Li J. Knowledge representation using interval-valued fuzzy formal concept lattice. Soft Computing, 2015, 19(1):1−18.

[9] Kengue JFD, Valtchev P, Djamegni CT. A parallel algorithm for lattice construction. In: Proc. of the Formal Concept Analysis. Berlin, Heidelberg: Springer-Verlag, 2005. 249−264.

[10] Krajca P, Outrata J, Vychodil V. Parallel recursive algorithm for FCA. In: Proc. of the CLA 2008. 2008. 71−82.

[11] Krajca P, Vychodil V. Distributed algorithm for computing formal concepts using map-reduce framework. In: Proc. of the Advances in Intelligent Data Analysis VIII. Berlin, Heidelberg: Springer-Verlag, 2009. 333−344.

[12] Dong H, Ma Y, Gong X. A new parallel algorithm for construction of concept lattice. Journal of Frontiers of Computer Science and Technology, 2008,2(6):651−657 (in Chinese with English abstract).

[13] Ma C. A parallel constructing algorithm based on dividing of closure system for concept lattice. China Management Informationization, 2009,12(21):20−24 (in Chinese with English abstract).

[14] Ma F, Zeng ZY, Yu JK. Research on vertically combine method of distributed concept lattices. Computer Engineering and Applications, 2011,47(34):68−63 (in Chinese with English abstract).

[15] Zhi HL. Extended model of formal concept analysis oriented for heterogeneous data analysis. Acta Electronica Sinica, 2013,41(12):2451−2455 (in Chinese with English abstract).

[16] Zhang Z, Chai YM, Wang LM, Fan M. A parallel algorithm generating fuzzy formal concepts. Pattern Recognition & Artificial Intelligence, 2013,26(3):260−269 (in Chinese with English abstract).

[17] Zhang Z, Du J, Wang LM. Load balance-based algorithm for parallelly generating fuzzy formal concepts. Control and Decision, 2014,29(11):1935−1942 (in Chinese with English abstract).

[18] Zhang T, Wei X, Hong W, Luan J. Attribute characteristics analysis and compare between AT and APOSD. In: Proc. of the 2015 5th Int'l Conf. on Instrumentation and Measurement, Computer, Communication and Control (IMCCC). IEEE, 2015. 947−951.

[19] Xu WH, Li JH, Wei L, Zhang T. Formal Concept Analysis: Theory and Application. Beijing: Science Press, 2016 (in Chinese).

[20] Zhang T, Wei X, Hong W, Li S. Transformation properties in attribute topology and attribute partial ordered structure diagram. In: Proc. of the 5th Int'l Conf. on Instrumentation and Measurement, Computer, Communication and Control (IMCCC). IEEE, 2015. 1085−1089.

[21] Zhang T, Li H, Wei X, Li L. Attribute topology and concept lattice bridged by concept tree. In: Proc. of the 5th Int'l Conf. on Instrumentation and Measurement, Computer, Communication and Control (IMCCC). IEEE, 2015. 1037−1041.

[22] Zhang T, Li H, Hong W, Yuan X, Wei X. Deep first formal concept search. Scientific World Journal, 2014,2014:275679−275679.

[23] Zhang T, Ren HL, Hong WX, Li H. The visualizing calculation of formal concept that based on the attribute topologies. Acta Electronica Sinica, 2014,42(5):925−932 (in Chinese with English abstract).

[24] Zhang T, Ren H, Wang X. A calculation of formal concept by attribute topology. ICIC Express Letters, Part B: Applications, An Int'l Journal of Research & Surveys, 2013,4:793−800. http://ci.nii.ac.jp/naid/40019602751

[25] Li G, Ma YC, Zhang T, Hong WX. Formal concept construction algorithm based on attribute topology. System Engineering—Theory & Practice, 2015,35(1):254−259 (in Chinese with English abstract).

[26] Wei L, Wan Q. Granular transformation and irreducible elements judgment based on pictorial diagrams. IEEE Trans. on Cybernetics, 2016,46(2):380−387.

[27] Hong WX, Li SX, Yu JP. A new approach of generation of structured partial ordered attribute diagram based on covering. ICIC Express Letters, Part B: Applications, 2015,6(4):1049−1054.

[28] Bai DH, Zhang T, Wei XY. Attributes-sorting algorithm based on attribute degree. Computer Engineering and Applications, 2015 (in Chinese with English abstract). http://www.cnki.net/kcms/detail/11.2127.TP.20150929.1045.018.html

[29] Carpineto C, Romano G. Concept Data Analysis: Theory and Applications. John Wiley & Sons, 2004.

[30] Gharehchopogh FS, Khaze SR. Data mining application for cyber space users tendency in blog writing: A case study. Int'l Journal of Computer Applications, 2013,47(18):40−46.

[31] Klahr D, Siegler RS. The representation of children's knowledge. Advances in Child Development and Behavior, 1978,12:62−116.

[32] Lincoff GH. The Audubon Society Field Guide to North American Mushrooms. Knopf: Distributed by Random House, 1981.

[33] Zupan B, Bohanec M, Bratko I, Demsar J. Machine learning by function decomposition. In: Proc. of the ICML. 1997. 421−429.

[34] Strok F, Neznanov A. Comparing and analyzing the computational complexity of FCA algorithms. In: Proc. of the 2010 Annual Research Conf. of the South African Institute of Computer Scientists and Information Technologists. ACM Press, 2010. 417−420.

[35] Kirchberg M, Leonardi E, Tan YS, Link S, Ko RKL, Lee BS. Formal concept discovery in semantic Web data. In: Proc. of the Formal Concept Analysis. Berlin, Heidelberg: Springer-Verlag, 2012. 164−179.

[36] Wray T, Eklund P. Using Formal Concept Analysis to Create Pathways through Museum Collections. Faculty of Engineering & Information Sciences Papers, 2014.

附中文参考文献:

[3] 孙小兵,李云,李必信,文万志.形式概念分析在软件维护中的应用综述.电子学报,2015,43(7):1399−1406.

[12] 董辉,马垣,宫玺.一种新的概念格并行构造算法.计算机科学与探索,2008,2(6):651−657.

[13] 马驰.基于闭包系统划分的概念格并行构造算法.中国管理信息化,2009,12(21):20−24.

[14] 马冯,曾志勇,余建坤.分布式概念格的纵向合并方法研究.计算机工程与应用,2011,47(34):68−71.

[15] 智慧来.面向异构数据分析的形式概念分析扩展模型.电子学报,2013,41(12):2451−2455.

[16] 张卓,柴玉梅,王黎明,范明.模糊形式概念并行构造算法.模式识别与人工智能,2013,26(3):260−269.

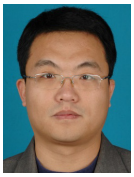
[17] 张卓,杜鹃,王黎明.基于负载均衡的模糊概念并行构造算法.控制与决策,2014,29(11):1935−1942.

[19] 徐伟华,李金海,魏玲,张涛.形式概念分析理论与应用.北京:科学出版社,2016.

[23] 张涛,任宏雷,洪文学,李慧.基于属性拓扑的可视化形式概念计算.电子学报,2014,42(5):925−932.

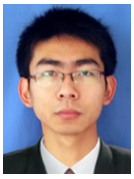
[25] 李刚,马彦超,张涛,洪文学.基于属性拓扑图的形式概念构造算法.系统工程理论与实践,2015,35(1):254−259.

[28] 白冬辉,张涛,魏昕宇.基于属性度的属性排序算法.计算机工程与应用,2015. http://www.cnki.net/kcms/detail/11.2127.TP.20150929.1045.018.html



张涛(1979－),男,河北唐山人,博士,副教授,CCF 专业会员,主要研究领域为形式概念分析,认知计算,医学信号处理.

李慧(1989－),女,硕士,主要研究领域为形式概念分析,属性拓扑.



白冬辉(1990－),男,硕士,主要研究领域为形式概念分析,认知计算.