

文章编号:1001-9081(2009)04-1110-04

交通流冗余数据识别和约简方法

王晓原, 吴芳, 邢丽

(山东理工大学 交通与车辆工程学院, 山东 淄博 255049)

(wangxiaoyuan@sdut.edu.cn)

摘要: 针对交通检测器检测到的数据存在冗余现象、影响后续决策并需要进行约简的问题, 提出了一种冗余数据的识别和约简方法。采用等级分组法实现对冗余数据的识别, 先通过等级法计算每个交通参数的权值并按照分组思想, 将大数据集分割成许多不相交的小数据集, 在各个小数据集中识别冗余数据。为避免漏查, 选择其他关键参数多次重复识别。识别出的冗余数据采用平均法约简。实例验证表明, 等级分组法识别冗余数据具有较好的精度, 随着阈值的增加, 查准率和查全率减小, 但仍在 93% 以上; 同时采用平均法约简, 拟合度较高, 达到 0.938。可见采用的冗余数据识别和约简方法能够有效地解决单数据源数据冗余问题。

关键词: 交通工程; 冗余数据; 等级法; 数据分组; 约简

中图分类号: TP182 **文献标志码:** A

Recognition and reduction of traffic flow redundant data

WANG Xiao-yuan, WU Fang, XING Li

(School of Transportation and Vehicle Engineering, Shandong University of Technology, Zibo Shandong 255049, China)

Abstract: The detected data often appear redundant, which affects the actual application of traffic models. A method of recognizing and reducing redundant data was proposed. Redundant data were recognized based on rank-based weights and packet method. Firstly, each of traffic parameters was endowed with certain weight according to rank-based weights method. Secondly, in terms of group thought, large data sets were divided into many non-intersecting small data sets. Finally, redundant data were detected and eliminated in each small data set. To avoid missing, the above steps can be repeated. And the recognized redundant data were reduced by average method. An application example shows that, the proposed recognition method of redundant data has a good detection precision, the recall and the precision decreased with the threshold increasing, but still over 93%. The reduced data have a high fitting degree, up to 0.938. The results indicate that, the problem of single data source can be solved effectively.

Key words: traffic engineering; redundant data; rank-based weights method; data packet; reduction

0 引言

智能运输系统 (Intelligent Transportation Systems, ITS) 建设被公认为是当前解决我国大城市交通拥挤和提高道路安全的有效手段。ITS 的实质是通过先进的技术, 收集交通参与者、车辆、道路系统和交管部门等各个交通要素的实时信息, 并使这些信息在上述要素之间有效流动, 从而强化它们的协调工作能力, 达到从整体上提高交通系统运作效率的目的。利用检测器收集交通信息的交通数据采集系统是 ITS 的基础子系统, 它为 ITS 其他系统的有效运行提供重要的数据支撑^[1-6]。ITS 各功能的顺利实现要求真实、准确、有效的交通流数据。理想情况下, 对于现实世界中的一个实体, 数据集中应该只有一条对应的记录^[7]。但由于检测器调试不正确、同一路段检测器布设过多等原因, 检测器输出的数据集极易存在信息冗余的问题, 造成动态交通流数据大幅增加, 不利于关键交通信息的凸显, 甚至可能导致建立错误的数据挖掘模型, 对后续的决策分析产生很大影响^[6,8], 因此, 需要对冗余数据进行约简。

就单检测器而言, 冗余数据主要是指检测器检测到的数

据构成集合中的相似重复数据。从狭义的角度来看, 检测器检测到的数据构成数据集, 如果其中两条记录在某些交通参数上的值相等或足够相似, 则认为这两条记录互为近似重^[7], 即存在冗余数据。对多检测器而言, 冗余数据主要是指, 为了获得城市交通系统运作的完整数据, 城市路网的相应路段需要布设检测器^[4], 但由于同一路段或相邻路段检测器布设密度过大, 直接影响了所采集的交通数据的有效性和准确性, 造成部分车辆信息重叠, 具有冗余性。限于篇幅, 本文对多检测器造成的冗余数据的约简方法不作研究, 仅研究单检测器冗余数据的约简方法。

目前, 对单检测器冗余数据, 即相似重复数据的研究多集中在文字信息的统计方面, 在交通领域, 对冗余数据的研究甚少。在文字信息统计方面, 主要采用距离函数模型^[9]、基于 q-gram 算法^[10]、“排序一合并”的方法^[7]等, 但上述方法在海量数据库中, 时间复杂度和空间复杂度较大, 并且排序时由于字符位置敏感性较高并不能保证相似的记录排在邻近的位置, 不能取得很好的效果。为了减少数据集中的冗余信息, 冗余记录的识别是一个关键步骤, 本文采用等级分组^[11-13]的方法, 根据等级法计算每个交通参数的权值及设计多趟查找方

收稿日期: 2008-10-07; 修回日期: 2008-12-05。

基金项目: 山东省自然科学基金资助项目(Y2006G32); 山东理工大学科研基金重点资助项目(2004KJZ02)。

作者简介: 王晓原(1970-), 男, 山东莱州人, 教授, 博士, 主要研究方向: 交通流理论与模拟、交通运输系统仿真、建模及优化; 吴芳(1982-), 女, 山东淄博人, 硕士研究生, 主要研究方向: 交通流理论与模拟; 邢丽(1982-), 女, 山东淄博人, 硕士研究生, 主要研究方向: 交通运输系统仿真、建模及优化。

法,提高了识别精度,同时采用分组法,降低了时间复杂度。

1 冗余数据识别和约简方法

1.1 基于等级分组法的冗余数据识别方法

1.1.1 基本定义^[11-12]

设数据集合 $X = \{x_1, x_2, \dots, x_n\}$, 交通参数向量 $F = \{F_1, F_2, \dots, F_p\}$, F_k 表示数据表第 k 个交通参数, 对于任意记录 $x_i = \{x_{i1}, x_{i2}, \dots, x_{ip}\}$, 其中 $1 \leq i \leq n$, x_{ip} 表示记录 x_i 第 p 维的值, 为叙述方便, 将日期、时间等也记为一交通参数。用 W_k 表示交通参数 F_k 的权值, 代表交通参数在对象中的重要程度, 称为交通参数的权重, 权重向量 $W = \{W_1, W_2, \dots, W_p\}$ 。

定义1 T_{ik} 是第 i 个操作用户为交通参数 F_k 所指定的等级(从1开始, 使用连续正整数表示等级, 1表示最高等级, 数值越大, 等级越低), T_k 表示第 k 个交通参数的最终统一等级, $k \in \{1, 2, \dots, p\}$, $i \in \{1, 2, \dots, N\}$, T_k 交通参数的最终统一等级:

$$T_k = \left\lfloor \left(\sum_{i=1}^N T_{ik} \right) / N \right\rfloor \quad (1)$$

定义2 采用RC(Rank-Centroid)转换方法^[12], 交通参数 F_k 的权重可以表示为:

$$W_k(RC) = \frac{1}{T} \sum_{t=T_k}^T \frac{1}{t} \quad (2)$$

T_k 表示 F_k 最终统一的等级, T 表示最低等级(即数值最大的等级), $k \in \{1, 2, \dots, p\}$ 。如果任意两参数的最终统一等级不相同, 那么 $T = p$, 如果存在两个或两个以上的交通参数, 它们的最终统一等级相同, 则式(2)应变成:

$$W_k = W_k(RC) / W' \quad (3)$$

定义3 对任意记录 x_i 与 x_j , 它们的第 k 维参数为 x_{ik} 与 x_{jk} , x_{ik} 与 x_{jk} 的相似度:

$$\text{SimField}(x_{ik}, x_{jk}) = \frac{\sum_{i=1}^q \max(score(a, x_{jk}))}{|x_{ik}|} \quad (4)$$

其中, $score(a, x_{jk})$ 表示 x_{ik} 中的数字 a 与 x_{jk} 中的每个数字匹配的分值, $0 \leq score(a, x_{jk}) \leq 1$, 如上述所定义; $|x_{ik}|$ 表示 x_{ik} 的长度; q 表示 x_{ik} 的数字的数量。

定义4 给定两条记录 x_i 和 x_j , 则 x_i 和 x_j 的记录相似度:

$$\text{SimRecord}(x_i, x_j) = \sum_{k=1}^p \text{SimField}(x_{ik}, x_{jk}) W_k \quad (5)$$

定义5 X_a 代表原数据集实际的重复记录集合, X_b 代表识别出来的重复记录集合, 查准率是正确识别出来的重复记录占识别出作为重复记录的比率, 则查准率表示为:

$$\text{ScanAccuracy}(X) = |X_a \cap X_b| / |X_b| \quad (6)$$

查全率是正确识别出来的重复记录占数据集中实际的重复记录比率, 则查全率表示为:

$$\text{ScanComplete}(X) = |X_a \cap X_b| / |X_a| \quad (7)$$

1.1.2 基本思想^[11-12]

1) 等级法计算权值。

本文采用RC等级转换法计算各交通参数的权重。等级法是一种计算各记录参数权重的方法, 其思想为:首先各用户根据实际经验为各个交通参数指定等级, 即最重要参数的等级指定为1, 第二重要的参数等级指定为2, 等; 然后根据式(1)计算各参数的最终统一等级; 最后根据式(2)或式(3)再计算它们相应的权重。表1为参数等级表。

2) 数据分组。

交通数据不断被检测得到, 构成海量数据库, 为提高冗余数据的识别效率, 需对大数据集作一定处理。根据分组思想,

把大的数据集分割成很多不相交的小数据集, 然后在各个小数据集中查找冗余数据, 为提高识别精度, 实行多趟查找。

基本思想:

①首先选择能明显区别记录间特征的交通参数, 把大数据集分割成很多个不相交的小数据集。不同领域数据集大小的判断标准不同, 就交通检测器检测到的记录数而言, 对检测器同一天的检测记录, 由于不同检测器采样间隔(30 s、2 min、5 min等)不同, 检测到的数据记录条数相应不同, 采样间隔越短, 记录条数越多, 数据集越大, 反之亦然; 另外, 采样间隔相同时, 采样时间长度不同, 数据集的大小不同, 时间长度越大, 数据集越大, 反之亦然。例如, 数据库中有若干天的数据, 可取日期作为分割依据, 把大数据集分割成数个不相交的集合。

②分割后, 若某些数据集仍然十分庞大, 则选择另外关键参数, 对这些数据集再次分割。如每天有24个小时, 构成的数据集仍然较大, 则对这些数据集进行二次分割, 取时间段, 把比较大的数据集再次分割成数个小数据集。

③若有些数据集仍很大, 可重复②, 直到数据集分割比较合理为止。另外, 引入多趟查找技术, 即把数据集划分成合理的小数据集, 并查找冗余记录, 这一轮结束后, 再选定另外关键参数或关键参数某些位, 重新对数据集进行划分, 并查找相似重复记录, 根据实际情况决定是否进行下一轮划分查找, 直至结果满意。

表1 交通参数等级表

交通参数	用户指定等级						等级
	U_1	U_2	...	U_i	...	U_N	
F_1	T_{11}	T_{21}	...	T_{11}	...	T_{N1}	T_1
F_2	T_{12}	T_{22}	...	T_{12}	...	T_{N2}	T_2
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
F_p	T_{1p}	T_{2p}	...	T_{1p}	...	T_{Np}	T_p

1.2 冗余数据的约简方法

对冗余数据的约简采用两种方法:当记录完全重复时, 删除多余重复记录, 只保留一条记录; 当记录相似时, 对流量、速度、占有率等各交通参数数据取平均值, 最终只含一条约简后记录。

2 应用实例

2.1 数据来源

本文所采集的数据是在其他外界干扰因素较少的情况下实测得到的, 采集地点为淄博市张店区张周路, 为简化计算, 本文所采用的数据集为同一天的数据, 采集时间是2008年5月21日7:00AM~19:00PM, 数据采样间隔为5 min, 应有数据记录144条。

原始流量数据本身并不存在冗余数据, 为了验证等级分组法识别冗余数据的有效性, 将原始数据进行改造, 在数据集的各交通参数中随机引入相似重复记录, 人为地制造一些冗余数据后, 数据记录变为154条。

2.2 算法流程

对冗余数据的约简事实上分两步进行, 主要是识别数据是否为冗余数据, 待确定冗余数据后再进行约简, 具体步骤如图1所示。

2.3 模型应用

对冗余数据的识别是在SQL Server 2000环境下进行的。由于本文采用的是同一天7:00 AM~19:00PM的数据, 故交通参数共有4个, F_1 、 F_2 、 F_3 、 F_4 分别表示时间、占有率、流量、速度。由于流量、占有率和速度都具有交通流数据的相同性

质,限于篇幅,以流量为例介绍冗余数据的识别和约简。

表 2 列出了部分含有人造冗余数据的原始数据。

表 2 部分含有人造冗余数据的原始数据

F_1	F_2	F_3	F_4	F_1	F_2	F_3	F_4	F_1	F_2	F_3	F_4
7:05	5	40	61	16:15	6	46	85
:	:	:	:	11:20	14	85	68	:	:	:	:
7:55	14	109	72	11:25	14	82	69	17:30	11	59	70
8:00	18	95	73	11:25	14	86	69	17:35	13	60	74
8:00	19	93	73	11:25	14	82	69	17:35	13	60	74
8:05	23	112	68	11:30	15	81	68	17:40	10	65	74
:	:	:	:	:	:	:	:	:	:	:	:
9:10	9	58	52	14:00	9	40	75	18:15	29	87	71
9:15	7	63	72	14:05	12	69	77	18:20	34	82	69
9:15	7	65	72	14:55	12	69	77	18:20	34	82	69
9:20	12	76	73	14:10	15	73	78	18:25	36	78	72
:	:	:	:	:	:	:	:	18:30	32	69	82
9:45	8	75	69	16:05	7	56	77	18:35	19	64	74
9:50	5	46	82	16:10	8	61	95	18:55	19	64	72
9:55	5	46	84	16:20	8	64	95	18:40	18	81	74

注:表中粗体部分表示人造冗余数据。

根据图 1 所示步骤,对检测器获得的数据集首先进行各参数权值的计算。随机选取负责交通数据处理的工作人员 10 名,每位工作人员根据自己的认识给出各参数的等级,然后计算各参数的权重,如表 3 所示。在得出权重的情况下,开始运行冗余数据的识别和约简。不同季节,不同地区,交通流数据的宏观规律不同,识别冗余数据所需的阈值也不尽相同。对阈值的合理选择,需要将数据集输入到 SQL Server 2000 软件中,反复调试阈值,直到输出的识别结果达到最佳。本文当阈值为 0.6 时,识别结果达到最佳,如图 2 所示,约简后的结果如图 3 所示,数据采样间隔为 5 min。

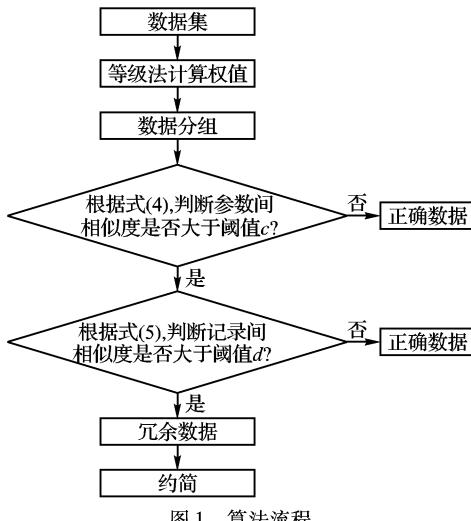


图 1 算法流程

表 3 等级表

交通参数	工作人员指定等级										等級	权重
	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}		
F_1	1	2	1	1	1	1	2	2	2	1	1.4	0.56399
F_2	2	1	2	3	2	1	1	1	1	2	1.6	0.21875
F_3	2	1	1	2	2	1	1	1	1	3	1.5	0.38542
F_4	3	2	3	3	2	1	1	3	1	3	2.2	0.06250

2.4 结果分析

2.4.1 查准率和查全率对比

完成冗余记录的识别后,和人工识别结果进行比较,根据式(6)、(7)计算查全率和查准率。为便于研究,实验数据集分为 3 组、2 组及不分组分别作为方法一、方法二、方法三,图

4 表示不同数据量下的查全率。从图 4 可以看出,随着数据量的增加,查全率增加。对数据集的准确分组,需要将数据集分为不同组后,分别计算查全率和查准率,当达到最优时,分组最合理。由于本文数据集较小,当数据集分为 2 组时,查全率和查准率达到最优,结果如表 4 所示。

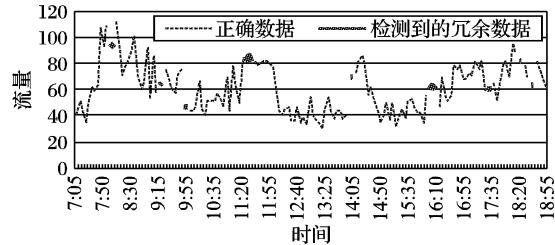


图 2 流量冗余数据识别结果

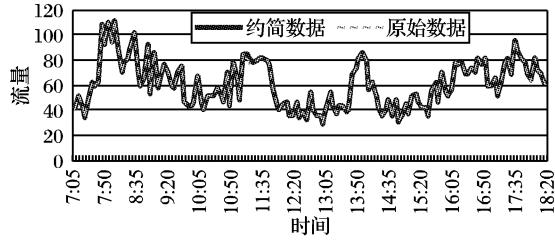


图 3 流量冗余数据约简后的结果与原始数据的比较

表 4 不同阈值下的识别结果比较

阈值 c	阈值 d	查全率 /%	查准率 /%
0.9	0.9	93.1	96.8
0.8	0.8	94.3	96.2
0.7	0.7	95.4	96.0
0.6	0.6	96.1	95.6

从图 2 和表 4 可以看出,等级分组法可以较好地识别出单检测器冗余数据,利用多趟分组查找可以有效地提高精度,随着阈值的减小,查全率增加,但查准率降低。

2.4.2 时间对比

为便于研究运行时间,使用 2.4.1 节分组方法,图 5 表示不同数据量下的运行时间。由于本文数据集较小,运行时间较快。三种方法的运行时间分别是 30 s、50 s、75 s。从图 5 可看出,随着数据量的增大,运行时间增加。

2.4.3 约简结果评价

为对约简结果进行评价,本文引入了拟合度(相关指数)

评价指标。计算公式分别为:

差异平方和:

$$Q = \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (8)$$

回归误差:

$$S = \sqrt{Q/(n-2)} \quad (9)$$

相关比:

$$RC = 1 - Q / \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (10)$$

其中 Q, S 的值愈小愈好,一般 $Q < 80$ 且 $S < 2.0$ 为宜;而拟合度 RC 的值越大越好,一般在 0.90 以上表示拟合较好。

约简结果的评价指标值如表 5 所示,结合图 5、表 5 可以看出,评价指标均满足要求,差异平方和较小,仅为 45,而拟合度达到了 0.956,可见,本文采用的约简方法具有较好的约简效果。

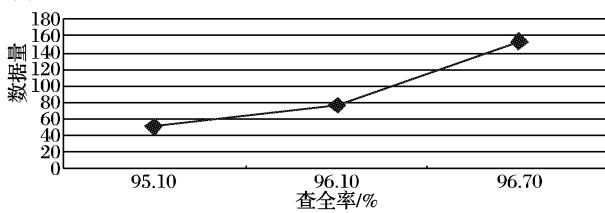


图 4 不同数据量下的查全率比较

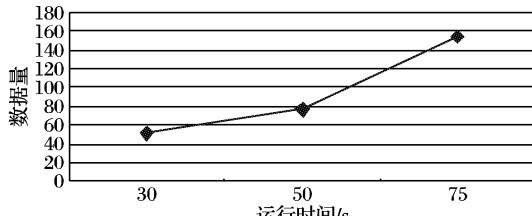


图 5 不同数据量下的运行时间比较

3 结语

本文采用等级法计算各交通参数的权重,对不同的参数使用不同的权重,从而提高了单检测器冗余数据的识别精度;采用分组法,有效地解决了大数据量的冗余数据识别问题,分组后在各个小数据集中进行冗余数据的识别,降低了时间复杂度,经过上述研究表明:等级分组法是一种识别单检测器冗

余数据的有效方法。由于“时间”这一交通参数占的权重较大,当其出现相似时,查准率较低,不易识别。另外,本文仅研究了单检测器冗余数据的识别,对多检测器冗余数据的识别需进一步研究。

表 5 冗余数据约简结果评价指标值

评价指标	结果
Q	45.000
S	1.412
RC	0.938

参考文献:

- [1] 王晓原, 隽志才, 贾洪飞, 等. 交通流突变分析的变点统计方法研究[J]. 中国公路学报, 2002, 15(4): 69–74.
- [2] 张敬磊, 王晓原. 交通事件自动检测算法研究进展[J]. 武汉理工大学学报: 交通科学与工程版, 2005, 29(2): 215–218.
- [3] 王晓原, 隽志才, 贾洪飞. 开发和评价 ITS 的微观交通流仿真模型[J]. 交通运输工程学报, 2002, 2(1): 64–66.
- [4] 伍建国, 王峰. 城市道路交通数据采集系统检测器优化布点研究[J]. 公路交通科技, 2004, 21(2): 88–91, 98.
- [5] 王晓原, 刘海红, 谭德荣. 交通流量变模式辨识的非参数概率变点模型[J]. 系统工程, 2006, 24(8): 19–22.
- [6] 姜桂艳. 道路交通状态判别技术与应用[M]. 北京: 人民交通出版社, 2004: 103–113.
- [7] 刘伟, 曹先彬. 对基于 MPN 的相似重复记录识别算法的改进[J]. 微计算机信息(管控一体化), 2005(24): 147–149.
- [8] 杨兆升. 基础交通信息融合技术及其应用[M]. 北京: 中国铁道出版社, 2005.
- [9] 黄健斌, 姬红兵, 孙鹤立. 近似重复记录的自适应距离度量检测[J]. 西安电子科技大学学报: 自然科学版, 2007, 34(2): 331–336.
- [10] 韩京宇, 徐立臻, 董逸生. 一种大数据量的相似记录检测方法[J]. 计算机研究与发展, 2005, 42(12): 2206–2212.
- [11] 李星毅, 包从剑, 施化吉. 数据仓库中的相似重复记录检测方法[J]. 电子科技大学学报, 2007, 36(6): 1273–1277.
- [12] 陈伟, 王昊, 朱文明. 一种提高相似重复记录检测精度的方法[J]. 计算机应用与软件, 2006, 23(10): 29–30, 42.
- [13] DEY D, SARKAR S, DE P. A distance-based approach to entity reconciliation in heterogeneous databases[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(3): 567–582.
- [14] 国科学: E 辑 信息科学, 2008, 38(2): 195–208.
- [15] ZHANG WEN-XIU, WEI LING, QI JIAN-JUN. Attribute reduction theory and approach to concept lattice[J]. Science in China Series F: Information Sciences, 2005, 48(6): 713–726.
- [16] GODIN R, MISSAOUI R, ALAOUI H. Incremental concept formation algorithms based on galois (concept) lattices[J]. Computational Intelligence, 1995, 11(2): 246–267.
- [17] GANTER B, KUZNETSOV S O. Stepwise construction of the Dedekind-MacNeille completion[C]// Proceedings of the 6th International Conference on Conceptual Structures: Theory, Tools and Applications. London: Springer-Verlag, 1998: 295–302.
- [18] LINDIG C. Fast concept analysis[C/OL]// STUMME G. Working with Conceptual Structures: Contributions to ICCS 2000. Aachen, Germany: Shaker Verlag, 2000: 152–161[2008–07–10]. <http://www.st.cs.uni-sb.de/~lindig/papers/lindig-fca-2000.pdf>.
- [19] 苗夺谦, 王国胤, 刘清, 等. 粒计算: 过去、现在与展望[M]. 北京: 科学出版社, 2007: 286–287.

(上接第 1109 页)

- [4] 商琳, 万琼, 姚望舒, 等. 一种连续值属性约简方法 ReCA[J]. 计算机研究与发展, 2005, 42(7): 1217–1224.
- [5] 刘清. 信息变换函数及动态信息系统[J]. 计算机科学, 2004, 31(10A): 15–17.
- [6] HE XIAO-WEI, XU LI-MING, SHEN WEN-ZHONG. Dynamic information system and its rough set model based on time sequence[C]// Proceedings of the 2006 IEEE International Conference on Granular Computing. Atlanta, Georgia, USA: [s. n.], 2006: 542–545.
- [7] WILLE R. Restructuring lattice theory: An approach based on hierarchies of concepts[C]// RIVALED I. Ordered Sets Ordered Sets. Boston: Reidel, 1982: 445–470.
- [8] GODIN R, MISSAOUI R. An incremental concept formation approach for learning from databases[J]. Theoretical Computer Science, 1994, 133(2): 387–419.
- [9] 王志海, 胡可云, 胡学钢, 等. 概念格上规则提取的一般算法与渐进式算法[J]. 计算机学报, 1999, 22(1): 66–70.
- [10] 仇国芳, 陈劲. 概念格的规则约简与属性特征[J]. 浙江大学学报: 理学版, 2007, 34(2): 158–162.
- [11] 魏玲, 邱建军, 张文修. 决策形式背景的概念格属性约简[J]. 中

- [12] 李鸿儒, 宋雪霞, 魏平等. 信息系统诱导出的形式背景及其性质[J]. 工程数学学报, 2005, 22(6): 970–974.
- [13] GODIN R, MISSAOUI R, ALAOUI H. Incremental concept formation algorithms based on galois (concept) lattices[J]. Computational Intelligence, 1995, 11(2): 246–267.
- [14] GANTER B, KUZNETSOV S O. Stepwise construction of the Dedekind-MacNeille completion[C]// Proceedings of the 6th International Conference on Conceptual Structures: Theory, Tools and Applications. London: Springer-Verlag, 1998: 295–302.
- [15] LINDIG C. Fast concept analysis[C/OL]// STUMME G. Working with Conceptual Structures: Contributions to ICCS 2000. Aachen, Germany: Shaker Verlag, 2000: 152–161[2008–07–10]. <http://www.st.cs.uni-sb.de/~lindig/papers/lindig-fca-2000.pdf>.
- [16] 苗夺谦, 王国胤, 刘清, 等. 粒计算: 过去、现在与展望[M]. 北京: 科学出版社, 2007: 286–287.