第41卷 第6A期
2014年6月

计 算 机 科 学
Computer Science

Vol. 41 No. 6A
June 2014

融合粗糙集和商空间的企业级信息系统日志挖掘方法

周丹晨

(中国工程物理研究院机械制造工艺研究所 绵阳 621900)

摘 要 为深度挖掘企业级信息系统用户群的多样化信息需求,通过宏观和微观的粒计算理论模型的对比分析,提出了一种融合粗糙集和商空间理论的企业级信息系统日志挖掘方法。首先以用户在一定时间内对企业级信息系统不同功能模块的使用频次和岗位角色来构建用户特征属性表;然后采用商空间理论的投影划分法进行用户群的层次化划分,得到两个不同用户粒度空间下的决策表;再利用基于粗糙集理论的知识获取方法,在两个用户粒度空间上分别导出相应的用户群识别规则;最终从不同角度综合分析用户群信息需求的一致性和差异性。应用实例验证了该方法的可行性和有效性。

关键词 粗糙集,商空间,粒计算,日志挖掘,企业级信息系统,用户粒度空间

中图法分类号 TP18 **文献标识码** A

Log Mining Method of Enterprise Information System by Combining Rough Set and Quotient Space

ZHOU Dan-chen

(Institute of Machinery Manufacturing Technology, China Academy of Engineering Physics, Mianyang 621900, China)

Abstract To realize the deep mining of diversification in information requirements among different user groups of enterprise information system, a log mining method by combining the theory of rough set and quotient space was put forward through comparison and analysis of the macrocosmic and microcosmic theoretical models of granular computing. Firstly, the users' characteristic attributes table was established in terms of their use frequencies for different function modules of enterprise information system in a certain period and their working roles. Secondly, the decision tables in two user granular spaces were built by hierarchical division of user groups based on projection division method of quotient space theory. Furthermore, the corresponding user identification rules in two user granular spaces were respectively derived by means of applying knowledge acquisition method of rough set theory. Finally, the consistency and discrepancy of information requirements among different user groups was comprehensively analyzed from different angles. The application case shows the feasibility and effectiveness of the proposed method.

Keywords Rough set, Quotient space, Granular computing, Log mining, Enterprise information system, User granular space

1 引言

企业级信息系统的用户涉及企业内不同部门、不同岗位、不同角色的各类工作人员,因此,能否满足不同用户群的多样化信息需求,为用户提供准确的、快捷的、个性化的信息服务,是企业级信息系统实施成败的关键要素之一。企业级信息系统日志挖掘的任务就是在系统运行后,通过对相当长一段时间内系统日志的挖掘和分析,全方位地发现用户最真实的信息需求,从而不断对系统予以完善和改进来满足用户的个性化需求,提高系统的运行效益。近年来,日志挖掘已成为数据挖掘领域重要的研究方向之一,但其研究几乎全部集中在Web日志挖掘方法上[1,2],而针对企业级信息系统日志挖掘方法的研究目前还未见报道。两者虽然都同样遵循数据挖掘的思路,挖掘过程大体相同,但在以下几个方面存在着较大的差异:

(1)挖掘目标不同。目前Web日志挖掘的目标主要是发现Internet用户在社会生活和电子商务方面的信息需求,而企业级信息系统日志挖掘的目标主要是发现企业Intranet用户的职业化信息需求。

(2)挖掘对象不同。Web日志挖掘的数据主要来源于Web服务器日志,多是半结构化或无结构化的数据,且存在多种多样的数据存储格式,而企业级信息系统日志挖掘是基于定制开发的基于数据库的结构化日志,格式标准统一。

(3)数据质量不同。与基于Internet的应用系统相比,企业级信息系统有更严格的用户管理和权限控制体系,因此其日志中的噪声数据比Web日志少得多,数据清理工作量较小。

(4)用户识别难度不同。用户识别是日志挖掘的重要步骤,也是Web日志挖掘的难点之一,这是由于Internet用户身份经常是隐性的、可随意更改的,再加上本地缓存、代理服

本文受中国工程物理研究院科学技术发展基金(2013B0203031)资助。

周丹晨(1969—),男,博士,高级工程师,硕士生导师,主要研究方向为数字化制造、人工智能,E-mail:zdc69@163.com。

务器、防火墙等的影响，造成很难认定用户身份的同一性，而在企业级信息系统日志挖掘中就相对清晰明了，因为系统的登录用户身份是显性的、预定义的、不可随意更改的。

因此，深入研究企业级信息系统日志挖掘方法对于拓展日志挖掘研究领域的广度、丰富其研究内涵具有积极的意义。

粒计算理论是目前人工智能研究领域的新热点，其研究主要包括粒的构造和基于粒的计算两个方面。数据挖掘的过程实际上是一个基于粒的计算过程，也就是在信息粒之间发现隐含的、有价值的关系的过程，因此粒计算理论为数据挖掘研究提供了一种新的途径，并日益显示出其强大的优势[3,4]。以不确定性处理为目标的粗糙集理论和以多粒度计算为目标的商空间理论，是两类粒计算理论模型的代表，如何通过两者的有机融合得到更有效的知识发现方法是当前粒计算研究的重要课题之一[5,6]。本文基于企业级信息系统日志，提出了一种通过粗糙集理论和商空间理论的融合应用，从不同的用户粒度空间挖掘和分析企业用户群信息需求的方法，为系统设计开发人员准确地获取与理解用户信息需求提供了一个新颖的技术手段。

2 两种粒计算理论模型的对比分析

粗糙集理论是一种处理模糊性和不精确性问题的数学工具[7-9]。其本质思想是利用等价关系（不可分辨关系）来建立论域的一个划分，得到不可区分的等价类，从而可以用精确的上近似集和下近似集来逼近一个边界模糊的集合，并通过对非重要属性的约简，得到精简的知识规则。

商空间理论模型可用一个三元组(X,f,Γ)来表示，其中X是论域，f是论域的属性，Γ是论域的结构，也就是X中各元素之间的相互关系（如拓扑关系）。在X上给定一个等价关系R，对应于R可以得到一个商集记为$[X]$，它对应的三元组$([X],[f],[\Gamma])$称为对应于R的商空间[10]。商空间理论的核心是构建合理的分层递阶结构，通过这种层次化结构的多重视角就可以从极不相同的粒度上观察和分析同一问题。

商空间理论和粗糙集理论都认为概念粒子可以用子集来表示，不同粒度下的粒子用不同大小的子集来描述，所有的粒子都通过等价关系获得划分产生[3]。但它们之间的区别主要在于：粗糙集理论的研究对象是由一个多值属性集合描述的一个对象集合，其论域是点集，各个对象之间没有结构关系或拓扑关系；而商空间理论是把商集作为描述不同粒度世界的数学模型，问题的不同粒度表示对应于不同的等价关系，即对论域进行不同的划分，这样就可以从不同粗细粒度的商空间来观察和分析同一个问题，以便得到对问题不同角度的理解，最终综合成对问题总的理解。总体来说，商空间理论可以看成是一种宏观的粒计算模型，而粗糙集理论则可以看成是一种微观的粒计算模型，即不同粒度均在同一个给定的空间结构中进行划分，没有粒度空间的变化。

3 融合粗糙集和商空间的企业级信息系统日志挖掘方法

3.1 学术思想

目前制造企业中主要有三大类人员，即技能人员（包括加工操作人员、质量检验人员、零件工时定额估算人员等子类人员）、技术人员（包括加工工艺编制人员、理化检测以及加工精度检测部门的技术人员等子类人员）和管理人员（包括生产车间的领导、生产车间的计划与调度人员以及企业生产管理人员等子类人员）。就每一大类人员来说，虽然其子类人员隶属于不同的部门，工作性质有所差别，信息需求必然存在差异性，但由于在岗位职责、工作关系以及绩效考核内容等方面具有一定的相似性，因此其信息需求同样也必然存在一些共同点。

基于以上分析，本文所提出的融合粗糙集和商空间的企业级信息系统日志挖掘方法的主要思想是：根据商空间理论的分层递阶结构思想进行论域的层次化划分，首先按用户所属大类进行用户群的划分，得到粗粒度的用户商空间$([X_1],[f_1],[\Gamma_1])$，在此用户粒度空间下，通过采用粗糙集理论的知识获取方法导出大类用户群识别规则，发现三大类用户群各自所具有的信息需求共同点；然后再按用户所属子类进行用户群的二次划分，得到细粒度的用户商空间$([X_2],[f_2],[\Gamma_2])$，并在此用户粒度空间下，同样通过采用粗糙集理论的知识获取方法导出相应的子类用户群识别规则，从而精细地刻画出各子类用户群之间细微的信息需求差异性。这样通过将宏观和微观的粒计算理论模型统一起来，从不同层次、不同角度所得到的融合信息，使用户群信息需求的分析结果具有了更全面、更直观、更易理解的实际价值。

3.2 流程和主要步骤

该方法的流程如图1所示。

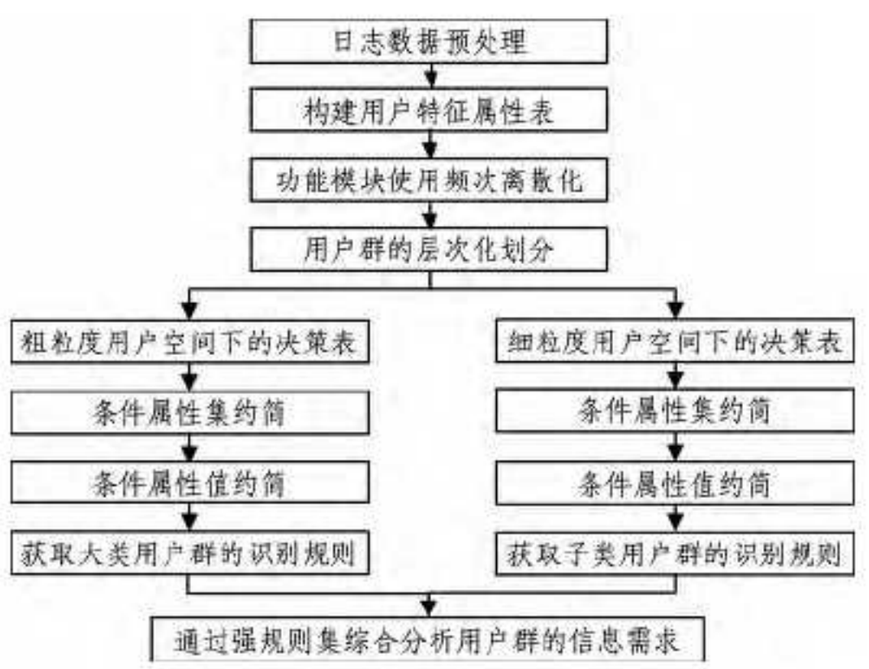

图1 融合粗糙集和商空间的企业级信息系统日志挖掘方法的流程

步骤1 设论域$U=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$是待分析的用户，由于企业级信息系统一般都是按照软件工程的模块化设计思想来开发的，不同的功能模块为企业用户提供不同类型、不同层次、不同展现方式的业务信息，以方便用户的操作，提高系统的使用效率，因此每一个用户由其在同样的时间段内对企业级信息系统不同功能模块的使用频次来表征，记作$x_i=\{y_{i1},y_{i2},\cdots,y_{ik},\cdots,y_{ih}\}$，其中$h$是系统功能模块的数量，$y_{ik}$是用户$x_i$对系统第$k$个功能模块的使用频次，从而构建起一个用户特征属性表。

步骤2 粗糙集理论只能分析离散型属性值，而用户的表征值即不同功能模块的使用频次是连续量，因此需要对数据进行离散化处理。

步骤3 按照商空间理论的投影划分法进行用户群的层

次化划分，得到两个不同用户粒度空间下的决策表，它们具有 h 个相同的条件属性，1 个分别按用户所属岗位大类和子类划分的决策属性。

步骤 4　对决策表进行属性约简，删去不重要的用户特征属性，从而在导出用户群识别规则时可以基于更简洁的条件。

步骤 5　对约简后的决策表进行属性值约简，去掉冗余的属性值，只保留不能去掉的核值，以简化用户群识别规则。

步骤 6　采用置信度和覆盖度两个关键指标对获取的用户群识别规则进行评价，通过强规则集综合分析用户群的信息需求。这两个指标分别表示了支持规则的实例在与规则具有相同条件属性的部分实例中的比例，以及在与规则具有相同决策属性的部分实例中的比例，突出反映了某一用户群识别规则的可信程度和在同一类用户群中的覆盖面大小。

4　应用实例与分析

某企业级制造信息集成平台于 2005～2006 年间设计开发并投入运行，主要由生产任务、技术信息、制造资源、工作成绩、数据统计与分析等 5 个功能模块组成[11]。基于平台的使用日志，采用本文所提出的方法，对不同用户群的信息需求进行了挖掘和分析。

在平台日志数据库记录的拥有全部功能模块使用权限的典型用户中，按技能人员、技术人员、管理人员三大类进行分组，然后根据用户所属子类从每组总使用频次位居前 50% 的人员中共随机选取了 24 位用户，分别统计从 2007 年 1 月 1 日至 2012 年 12 月 31 日各个功能模块的使用频次，从而构建用户特征属性表，如表 1 所列。

表 1　用户特征属性表

用户	功能模块使用频次					用户岗位角色	
	生产任务	技术信息	制造资源	工作成绩	数据统计与分析	所属大类	所属子类
x_1	8171	51	52	7425	168	技能人员	加工操作人员
x_2	5409	36	23	6732	75	技能人员	加工操作人员
x_3	4365	5	5	7196	88	技能人员	加工操作人员
x_4	64	173	138	19568	88	技能人员	工时估算人员
x_5	68	64	28	10382	165	技能人员	工时估算人员
x_6	12	24	113	6902	14	技能人员	工时估算人员
x_7	8	148	18	3116	20	技能人员	质量检验人员
x_8	4	249	1	1492	13	技能人员	质量检验人员
x_9	11	170	27	1159	19	技能人员	质量检验人员
x_{10}	72	1165	1389	430	5	技术人员	加工工艺编制人员
x_{11}	6	550	738	615	59	技术人员	加工工艺编制人员
x_{12}	13	963	875	555	6	技术人员	加工工艺编制人员
x_{13}	19	1815	21	4	11	技术人员	产品检测技术人员
x_{14}	58	940	64	32	67	技术人员	产品检测技术人员
x_{15}	1	815	38	1	17	技术人员	产品检测技术人员
x_{16}	97	77	76	336	4894	管理人员	生产车间领导
x_{17}	73	25	16	240	2700	管理人员	生产车间领导
x_{18}	63	33	54	259	2387	管理人员	生产车间领导
x_{19}	65	41	31	1644	5899	管理人员	车间计划调度人员
x_{20}	98	53	56	1281	4451	管理人员	车间计划调度人员
x_{21}	86	63	47	1352	4035	管理人员	车间计划调度人员
x_{22}	387	371	89	84	412	管理人员	企业生产管理人员
x_{23}	339	282	117	82	550	管理人员	企业生产管理人员
x_{24}	361	296	123	104	517	管理人员	企业生产管理人员

采用式(1)进行功能模块使用频次的离散化处理：

$$y'_{ik}=\frac{y_{ik}-\bar{y}_k}{\bar{y}_k}\times 100\% \tag{1}$$

其中，$\bar{y}_k$ 为平台日志中所记录的所有用户对第 k 个功能模块的人均使用频次。根据 y'_{ik} 值的范围，对照表 2 即可将 y_{ik} 转换为相应的离散值，从而由表 1 得到两个不同用户粒度空间下的决策表，如表 2 所列。

表 2　功能模块使用频次的离散区间

y'_{ik}	≤−50%	(−50%, −10%)	[−10%, 10%]	(+10%, 50%)	≥50%
离散值	1	2	3	4	5
含义	很低	较低	一般	较高	很高

4.1　发现三大类用户群各自信息需求的共同点

综合采用代数法和基于条件信息熵的 CEBARKCC 算法进行粗粒度用户空间下决策表的属性约简[12]。

首先用代数法确定决策表的核属性集。计算可得：$POS_{\{C-a_1\}}(D)=POS_{\{C-a_2\}}(D)=POS_{\{C-a_3\}}(D)=POS_{\{C-a_4\}}(D)=POS_{\{C-a_5\}}(D)=POS_C(D)$，因此，粗粒度用户决策表的核属性集为空，即 $CORE_D(C)=\varnothing$；然后逐次选择条件信息熵值为最小值的条件属性添加到约简集中，直至满足指定的终止条件。第一步计算可得：$H(D|C)=0$，$H(D|a_1)=0.358$，$H(D|a_2)=0.224$，$H(D|a_3)=0.363$，$H(D|a_4)=0.290$，$H(D|a_5)=0.183$，由于 $H(D|a_5)$ 最小，因此首先选择属性 a_5 作为约简集中的条件属性，且因为 $H(D|a_5)\neq H(D|C)$，因此约简过程以 a_5 为基继续进行。第二步计算可得：$H(D|\{a_1,a_5\})=0.151$，$H(D|\{a_2,a_5\})=0$，$H(D|\{a_3,a_5\})=0.111$，$H(D|\{a_4,a_5\})=0$，由于 $H(D|C)=H(D|\{a_2,a_5\})=H(D|\{a_4,a_5\})$，因此属性约简过程结束，得出相对约简集 $RED_D(C)=\{a_2,a_4,a_5\}$。

表 3　两个不同用户粒度空间下的决策表

U	a_1	a_2	a_3	a_4	a_5	D	D′
x_1	5	1	1	5	1	Ⅰ	$Ⅰ_1$
x_2	5	1	1	5	1	Ⅰ	$Ⅰ_1$
x_3	5	1	1	5	1	Ⅰ	$Ⅰ_1$
x_4	1	2	2	5	1	Ⅰ	$Ⅰ_2$
x_5	1	1	1	5	1	Ⅰ	$Ⅰ_2$
x_6	1	1	2	5	1	Ⅰ	$Ⅰ_2$
x_7	1	2	1	5	1	Ⅰ	$Ⅰ_3$
x_8	1	4	1	5	1	Ⅰ	$Ⅰ_3$
x_9	1	2	1	5	1	Ⅰ	$Ⅰ_3$
x_{10}	1	5	5	2	1	Ⅱ	$Ⅱ_1$
x_{11}	1	5	5	4	1	Ⅱ	$Ⅱ_1$
x_{12}	1	5	5	3	1	Ⅱ	$Ⅱ_1$
x_{13}	1	5	1	1	1	Ⅱ	$Ⅱ_2$
x_{14}	1	5	1	1	1	Ⅱ	$Ⅱ_2$
x_{15}	1	5	1	1	1	Ⅱ	$Ⅱ_2$
x_{16}	1	1	1	2	5	Ⅲ	$Ⅲ_1$
x_{17}	1	1	1	1	5	Ⅲ	$Ⅲ_1$
x_{18}	1	1	1	1	5	Ⅲ	$Ⅲ_1$
x_{19}	1	1	1	5	5	Ⅲ	$Ⅲ_2$
x_{20}	1	1	1	5	5	Ⅲ	$Ⅲ_2$
x_{21}	1	1	1	5	5	Ⅲ	$Ⅲ_2$
x_{22}	4	5	2	1	4	Ⅲ	$Ⅲ_3$
x_{23}	4	4	2	1	5	Ⅲ	$Ⅲ_3$
x_{24}	4	4	2	1	5	Ⅲ	$Ⅲ_3$

采用数据分析法进行属性值约简，分析删除了一个条件属性值后是否改变了决策表的不可分辨关系，最终得到仅包含属性核值的用户决策表，从中获取具有高置信度和覆盖度的用户群识别规则集，如表 4 所列。从中发现三大类用户群各自信息需求的共同点：技能人员对能细致反映个人工作量

的信息十分关注，技术人员对工艺规程、理化检测报告、精密检测报告等技术类信息的需求强烈，但两者均对不同车间、不同班组、不同工种之间的工作负荷对比以及变化趋势等较为宏观的信息没有兴趣，而这些信息却正是管理人员非常关心的。

表 4　三大类用户群的强识别规则集

编号	大类识别规则	置信度	覆盖度
R_1	$a_4=5$ and $a_5=1 \rightarrow D=\mathrm{I}$	9/9=1	9/9=1
R_2	$a_2=5$ and $a_5=1 \rightarrow D=\mathrm{II}$	6/6=1	6/6=1
R_3	$a_5=5 \rightarrow D=\mathrm{III}$	8/8=1	8/9=0.889

4.2　发现子类用户群信息需求的差异性

采用同样的计算步骤，可以在大类用户群识别规则的前提下，从细粒度用户空间下的决策表中获取各子类用户群的强识别规则集，如表 5 所列。

表 5　各子类用户群的强识别规则集

编号	大类识别规则	子类识别规则	置信度	覆盖度
R_1'	R_1	$a_1=5 \rightarrow D'=\mathrm{I}_1$	3/3=1	3/3=1
R_2'		$a_1=1 \rightarrow D'=\mathrm{I}_2$ or $D'=\mathrm{I}_3$	6/6=1	6/6=1
R_3'	R_2	$a_3=5 \rightarrow D'=\mathrm{II}_1$	3/3=1	3/3=1
R_4'		$a_3=1 \rightarrow D'=\mathrm{II}_2$	3/3=1	3/3=1
R_5'	R_3	$a_1=1$ and $a_4=1 \rightarrow D'=\mathrm{III}_1$	2/2=1	2/3=0.667
R_6'		$a_1=1$ and $a_4=5 \rightarrow D'=\mathrm{III}_2$	3/3=1	3/3=1
R_7'		$a_1=4$ and $a_4=1 \rightarrow D'=\mathrm{III}_3$	3/3=1	3/3=1

从表 5 中可以发现各子类用户群信息需求的差异性：对于三类技能人员来说，其差异性集中体现在对生产任务类信息的需求上，可以看出加工操作人员非常关心车间生产任务的进展情况，迫切希望了解未来一段时间本人和所在班组将要承担的加工任务；对于两类技术人员来说，其差异性突出显现在对制造资源类信息的需求上，可以发现工艺编制人员在此方面的信息需求强烈，说明他们必须通过掌握加工设备的工艺参数与运行状态以及车间刀具、工装等库存信息来确定合理可行的工艺路线，同时也需要深入了解企业器材材料的库存情况，以便准确地编制工艺规程的材料清单；三类管理人员的信息需求相比较而言，企业生产管理人员作为整个企业生产任务的全局性计划和协调人员，比其他两者更关注生产任务类信息，而车间计划调度人员却更关心车间个人工作量方面的信息。

结束语　满足用户的多样化信息需求是企业级信息系统实施成功与否的关键要素之一。基于粗糙集和商空间理论的思想，本文提出了一种通过企业级信息系统的日志挖掘，从不同的用户粒度空间分析企业用户群信息需求一致性和差异性的技术方法，并通过一个运行多年的企业级信息系统的应用实例分析，验证了方法的可行性和有效性。如何从单个功能模块的使用中发现更深层次的用户信息需求，以及如何利用信息推送技术将获取的用户客观信息需求转化为信息系统细致入微的主动信息服务，将是下一步研究工作的重点。

参 考 文 献

[1] 付博，赵世奇，刘挺. Web 查询日志研究综述[J]. 电子学报，2013，40(9)：1800-1808

[2] 何跃，马丽霞，腾格尔. 基于用户访问兴趣的 Web 日志挖掘[J]. 系统工程理论与实践，2012，32(6)：1353-1361

[3] 王国胤，张清华，胡军. 粒计算研究综述[J]. 智能系统学报，2007，2(6)：8-26

[4] 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

[5] 张钹，张铃. 粒计算未来发展方向探讨[J]. 重庆邮电大学学报：自然科学版，2010，22(5)：538-540

[6] 薛志远，张清华. 复合粒计算模型研究进展[J]. 重庆邮电大学学报：自然科学版，2010，22(5)：631-640

[7] 王国胤，姚一豫，于洪. 粗糙集理论与应用研究综述[J]. 计算机学报，2009，32(7)：1229-1246

[8] 苗夺谦，李道国. 粗糙集理论、算法与应用[M]. 北京：清华大学出版社，2008

[9] 吴伟志，杨玉芳. 空间遥感数据的多粒度标记分类[J]. 计算机科学，2012，39(4)：23-27

[10] 张铃，张钹. 问题求解理论及应用_商空间粒度计算理论及应用[M]. 北京：清华大学出版社，2007

[11] 周丹晨，尚鎏，周战强. 面向产品制造全过程的企业信息集成平台研究[J]. 计算机应用与软件，2008，25(7)：149-151

[12] 王国胤，于洪，杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报，2002，25(7)：759-766

(上接第 416 页)

如图 6，因为时间序列和空间插值的结合，使得插值误差有了很大的降低，平均在 1℃ 左右，这对插值精度的提升具有很大的意义。

结束语　基于时间序列模型时空插值算法，加入时间序列模型的插值算法因为可以既考虑空间因素的影响又同时考虑时间因素对气象数据的影响，避免了空间插值的单一性。实验表明，它比目前经常使用的数据修补方法更加精确，不仅实现了已经残缺的气象观测序列的修补，其最大的贡献在于解决了极端天气下线路故障或气象设备损坏而收不到气象观测站的实时气象数据的问题，方便气象敏感部门（如电力部门）根据实时的情况做出相应的措施。但是由于数据的变化不一定是严格按某一规律进行的，因此该方法不能保证对所有的时间序列都适合，尤其对于数据突变明显的时间序列。

参 考 文 献

[1] 矫梅燕. 在改革中探索精细化预报业务发展[N]. 中国气象报，2006，10(26)

[2] 刘晓晓. 气象监测数据的时空特征分析与建模研究[D]. 开封：河南大学，2009

[3] 李莎，舒红，等. 利用时空 kriging 进行气温插值研究[J]. 武汉大学学报，2012(2)：237-240

[4] 张建龙，王玲. 重庆岩溶区的气候时空变化特征分析及趋势预测[J]. 热带气象学报，2008(6)：239-248

[5] 徐爱萍，胡力，等. 空间克里金插值的时空扩展与实现[J]. 计算机应用，2011(31)：273-276

[6] 彭思岭. 气象要素时空插值方法研究[D]. 长沙：中南大学，2010

[7] 谢福鼎，王赫楠，等. 基于函数的时间序列分段线性表示方法[J]. 计算机科学，2011，38(11)：153-155