

集成共享地理空间数据格网化编码研究

刘 鹏^{1,2} 王卷乐² 陈二洋^{1,2} 卜 坤³

(1. 中国矿业大学(北京), 地球科学与测绘工程学院, 北京 100083; 2. 中国科学院地理科学与资源研究所, 资源环境信息系统国家重点实验室, 北京 100101; 3. 中国科学院东北地理与农业生态研究所, 吉林长春 130102)

摘要: 针对陆地表层地理空间数据的集成共享问题, 提出从数据分类、数据编码到数据格网化编码的解决方案。从我国国情出发, 以中国国家地理格网为基础, 分析其空间数据的组织与编码体系, 并针对国家科技基础条件平台—地球系统科学数据共享平台中的陆地表层数据, 提出结合数据分类的数据属性编码, 进而建立能够与国家地理格网相衔接的地理空间数据格网化编码应用方案。以地理空间数据的点、面实体数据为例, 证实了该编码的应用可行性。预期该编码方案能够在科学数据资源管理、数据检索访问、数据资源集成、数据资源格网化处理、数据资源空间分析、数据标识与引用等方面发挥作用。

关键词: 地理格网; 数据编码; 地球系统科学; 地理空间数据; 数据共享

中图分类号: TP399

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2014.04.012

Research on Geospatial Data Grid Coding for Data Integration and Sharing

Liu Peng^{1,2} Wang Juanle² Chen Eryang^{1,2} Bu Kun³

(1. China University of Mining & Technology, College of Geoscience and Surveying Engineering, Beijing 100083; 2. Institute of Geographic Sciences and Natural Resources Research Chinese Academy of Sciences, State Key Laboratory of Resources and Environmental Information System, Beijing 100101; 3. Northeast Institute of Geography and Agroecology Chinese Academy of Sciences, Changchun 130102)

Abstract: Grid is not only the special reference base, but also the container of multi sources geospatial data. This paper discussed the problem of the current geospatial data integration, which pointed out that the geographic grid was an effective method to solve the problem. According to China national situation, the national geographic grid organization and coding system was analyzed firstly; land surface data coding schema was designed based on the data classification system of Earth System Science Data Sharing Platform; then an integration solution for land surface grid coding is proposed. Actual point and polygon data resources coding cases in Earth System Science Data Sharing Platform were experimented in the study. Results showed that the grid coding solution is feasible for geospatial integration and sharing. It can be anticipated that the grid coding of geospatial data will play important roles on data resources management, access, integration, grid processing,

作者简介: 刘鹏 (1991-), 男, 硕士研究生, 研究方向: 地理信息系统开发; 王卷乐* (1976-), 男, 副研究员, 研究方向: 资源环境信息集成与共享; 陈二洋 (1991-), 男, 硕士研究生, 研究方向: 遥感与地理信息系统; 卜坤 (1980-), 男, 博士, 研究方向: 遥感与GIS应用研究

基金项目: 科技基础性工作专项项目“格网化资源环境综合科学调查规范”(2011FY110400); 中国科学院信息化专项项目“资源学科领域基础科学数据整合与集成应用”(XXH12504-1-01)。

收稿日期: 2014年7月12日。

spatial analysis, identification and citation.

Keywords: geographic grid, data coding, earth system science, geospatial data, data sharing

1 引言

伴随着科技的迅速发展,人类对地球的测量手段不断进步,以全球定位系统GPS为代表的单点定位和以遥感卫星为代表的平面测量得到了广泛的应用。测量手段的进步和海量数据的快速积累为数据集成提出了新的、更高的挑战。以往的矢量、栅格数据模型难以满足集成越来越多的多源现势空间数据与历史空间数据的集成需求^[1]。

地理格网的数据管理模式为这一问题的解决提供了契机。格网化数据集成不但可以提高空间信息集成的效率,还可以减小数据精度损失和资源消耗^[2]。格网模型有广义与狭义之分,本文所述的格网模型为狭义格网模型,即以正方形为基本格网单元的地理格网模型。格网模型在国外已有广泛应用,例如美国正在建设的重要基础设施全球信息栅格(GIG)是全球互联、终端到终端的信息集合,用于信息的收集、处理、存储、传播和管理,为作战人员、决策者等提供技术支持^[3]。美国国防信息系统局先后分别于2006、2011、2012年发布了第一代、第二代、第三代《“全球信息栅格”整合主计划》,提出了利用云计算技术手段,结合高频传输卫星技术,为军事决策提供有力支持,将信息优势变成决策优势。此外,国外的数据共享网站如国际农业研究咨询顾问集团空间信息协会(Consultative Group for International Agriculture Research – Consortium for Spatial Information, CGIAR–CSI)通过全球地理格网发布航天飞机雷达地形测量(SRTM)数据^[4]。国内格网模型的综合集成性应用较少,更多的是学者对社会经济、土地利用等专题数据进行格网化研究。如廖顺宝等根据海拔高度、土地利用、主要道路分布等信息制成1km格网人口密度分布图^[5];张超等基于格网数据对日本新川破堤洪水泛滥进行了模拟,并对灾害损失进行了评

估^[6];丁志雄等基于GIS格网模型对洪涝灾害损失进行评估研究^[7];刘明亮等由全国1:10万比例尺资源环境遥感数据库生成1km格网土地利用/土地覆盖数据^[8]。

数据编码是指在数据分类的基础上,赋予数据一定规律性、计算机容易识别与处理的符号,其作用在于保证科学数据的有效管理,并能支持高效率的查询、集成处理服务和共享利用^[9]。引入地理格网思想的地理空间数据编码,有利于实现多源地理空间数据的集成共享。本文根据我国国情,以国家科技基础条件平台——地球系统科学数据共享平台中的数据资源为对象,提出以中国国家地理格网为基础,融合数据分类、数据编码、格网化编码应用的系统解决方案,以期为地理空间数据的集成和共享利用提供参考。

2 国家地理格网编码

《国家地理格网标准》(GB/T 12409–2009)规定了地理格网的划分与编码方法,为各类地理空间数据的整合提供了空间参照。该标准定义了经纬坐标格网和直角坐标格网两种格网划分方法,前者适合大范围(全球、全国)空间信息应用,后者适用于小范围(省区、城乡)精确分布应用^[2]。

经纬坐标格网按经纬度差分级,以1″经纬度差作为基本单元。经纬坐标格网分为5级,见表1。

经纬格网坐标代码由5类元素组成:象限代码(东、西、南、北半球代码)、格网间隔代码、间隔单位代码、经纬度代码和格网代码(图1)。

表1 经纬坐标格网分级

格网间隔	1°	10′	1′	10″	1″
格网名称	一度格网	十分格网	分格网	十秒格网	秒格网

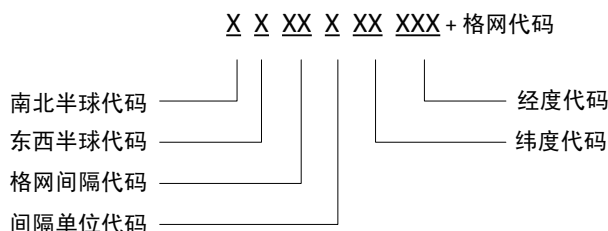


图1 经纬坐标格网编码结构图

前4类元素为必选元素，格网代码为根据具体情况而定的可选代码。

该编码方案对任何一个地理空间位置而言，都能得到与之相对应的格网编码。然而，选取不同的格网间隔单元，其编码结果是不一致的。以中国科学院生态系统网络盐亭野外台站（其地理位置为东经105度28分，北纬31度16分）为例，其经纬度格网的编码如表2所示。

表2 盐亭站不同格网间隔编码

格网间隔	格网编码
一度格网	NE01D31105
十分格网	NE10M311050102
一分格网	NE01M311051628
十秒格网	NE10S3110516280000
一秒格网	NE01S3110516280000

上述编码方案的优点是实现了多种格网尺度下的唯一格网编码，但其存在的问题有两点。一是编码长度不定长，不同格网尺度下的同一位置的编码长度不同，例如表2中一度格网的编码长度是10位，一秒格网的编码长度是18位。二是格网尺度跳跃，例如十分格网是一分格网单元的10倍，而一度格网是十分格网单元的6倍，这在未来利用计算机进行格网化数据管理时会出现编码困难、索引效率低下等问题。尽管有这样的问題，但其格网化编码思想仍可为地理空间数据集成管理中的数据编码提供基础。

3 陆地表层数据格网化编码方案

3.1 格网化编码思路

国家科技基础条件平台——地球系统科学数

据共享平台于2003年作为国家科学数据共享工程首批试点之一启动^[10]，其主要目标是整合、集成科研院所、高等院校和科学家个人通过科研活动所产生的研究型分散科学数据^[11]。王卷乐等借鉴国际大型数据共享机构的数据分类思想，结合平台的服务实践，提出了既具有我国特色又能满足该平台用户需求和数据集成要求的数据分类方案^[12]。其中一级类包括：大气圈、陆地表层、陆地水圈、冰冻圈、自然资源、海洋极地、固体地球、古环境、日地空间环境与天文、遥感数据。本文研究对象陆地表层的二级类包括：基础地理、土地利用/覆盖、人口、社会经济、区划、地形、地貌、土壤、沙漠/荒漠、湖泊/水库、湿地、植被、生态系统、环境、灾害。

当前的国家地理格网标准中的空间编码方案可以对所有的地理空间数据进行编码，并且保证所有的地理空间数据遵从统一的空间参考基准。现有的地球系统科学数据共享平台也已建立了系统的数据分类方案，且能保证所有陆地表层的地理空间数据能够有可管理的数据目录。如果能够将地理格网编码与数据分类编码衔接起来，则可以建立既适合于格网化数据管理又能满足科学数据集成管理的编码体系。本研究因此提出“格网编码”+“数据分类编码”的陆地表层地理空间数据编码思路。

3.2 数据编码

陆地表层数据格网化编码方案包括两段编码。第一段为国家地理格网编码，第二段为数据分类编码。如图2所示，格网编码直接采用《国家地理格网标准 GB/T 12409-2009》编码方法。数据编码段采用11-17位编码结构，首位采用1

位标识码——G表示地球系统科学数据；一级类码对应地球系统科学数据分类体系中一级类编码，采用1位字母表示；二级类码对应分类体系中的二级类编码，采用2位数字表示；属性码采用6位数字表示，从左到右分别表示数据的比例尺、时间分辨率、空间尺度、空间分辨率、类型、产生方式；顺序码用5位数字表示，保证数据编码唯一性；版本号用2位数字表示，便于数据更新记录。

在该编码结构中，属性码的编码要求最为复杂，有必要特别进行说明。属性码段共6位，第一位是数据比例尺，分为1:400万以上、1:400万、1:100万、1:25万、1:10万、1:5万、1:5万以下；第二位是数据时间分辨率，分为天以下、天、月/季、年、十年、百年及以上；第三位是空间尺度，分为全球尺度、洲际尺度、国家尺度、省级尺度、市级尺度、县级尺度、县级以下；第四位是空间分辨率，分为十公里级及以上、公里级、百米级、十米级、米级及以下；第五位是数据类型，分为矢量、栅格（非遥感影像）、栅格（遥感影像）、属性（表格）、图片、视频；第六位是数据产生方式，分为站点观测、实验试验、计算模拟、遥感解译、遥感反演、考察调查、统计资料、网络挖掘、数字化加工等。具体编码规则按上述顺序由小到大进行数字编码，不在上述类型中则编码为0。

属性码直接刻画科学数据的具体特征，这是数据共享中用户最关心的信息之一。然而，该部分编码对数据的更新状态非常敏感，其编码值可能会因数据状态的变化而变化。因此，在实际应用中，根据需求的不同建议将此部分列为可选编码段。例如，针对用户服务，建议提供包含属性

码的数据编码信息，便于用户检索和了解数据内容，这种情况可称为长码；针对数据的标识和管理，建议提供不包含属性码的数据编码信息，便于数据的永久标识和存储管理，这种情况可称为短码。

4 应用案例

陆地表层数据主要包括矢量、栅格、属性数据，其中矢量数据分为点状、线状、面状数据，考虑到线状数据如河流等其数据实体往往呈面状分布，故可将线状数据视为面状数据。本文各选取一个点数据和一个面数据为例进行应用试验。

4.1 点数据编码应用

利用地球系统科学数据共享网《盐亭站1985—1994年旱地水分、养分动态观测数据集》，对点状数据编码进行试验^[13]。该数据的地理位置为东经105度28分，北纬31度16分，选取1'格网间隔，根据《国家地理格网标准GB/T 12409—2009》，其经纬度格网编码为：NE01M311051628。在地球系统科学数据分类体系中，该数据所在二级类生态系统编码为13，其所属一级类陆地表层编码为L。该数据为观测数据，无比例尺编码为0；该数据为逐年观测，时间分辨率为年，编码为4；空间尺度为县级以下，编码为7；数据无空间分辨率，其编码为0；数据类型为属性表格，编码为4；数据生产方式为观测数据，编码为1。为该数据取顺序码为00201，版本号为01。其数据编码为：GL130470410020101。最终形成该数据的格网化数据编码为：NE01M311051628GL130470410020101，短码：NE01M311051628GL130020101。以上编码实现了地球系统科学数据共享网中一级和二级分类的整合，能够完整表

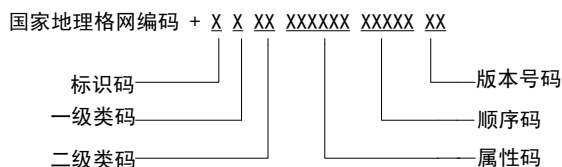


图2 地球系统科学数据码位结构图

达数据的分类信息。

4.2 面数据应用

面状数据往往涉及多个格网单元编码。在明确面状数据的空间范围前提下，可通过一致性的格网编码，实现面状数据的无缝表达。如图3中该数据涉及的格网编码分别为：

NE01M431122724, NE01M431122725,
NE01M431122726, NE01M431122624,
NE01M431122625, NE01M431122626,
NE01M431122524, NE01M431122525。

图4为“攀西地区滑坡、泥石流、崩塌分布以及危险度区划数据”实例，其对应的数据编码分别是：

GL154050190220101, GL154050190220201,

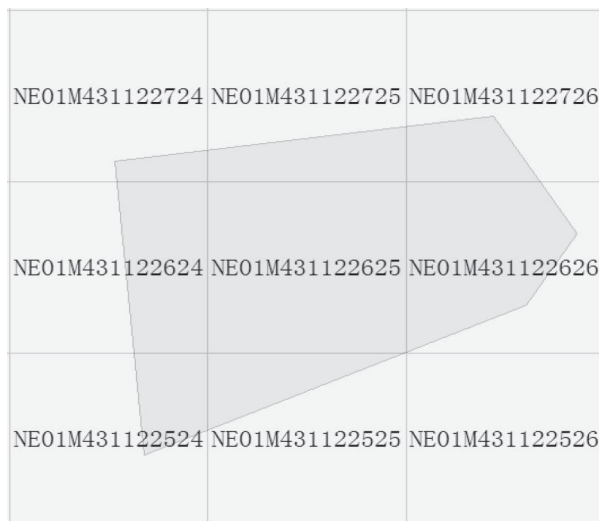


图3 面状矢量数据格网编码

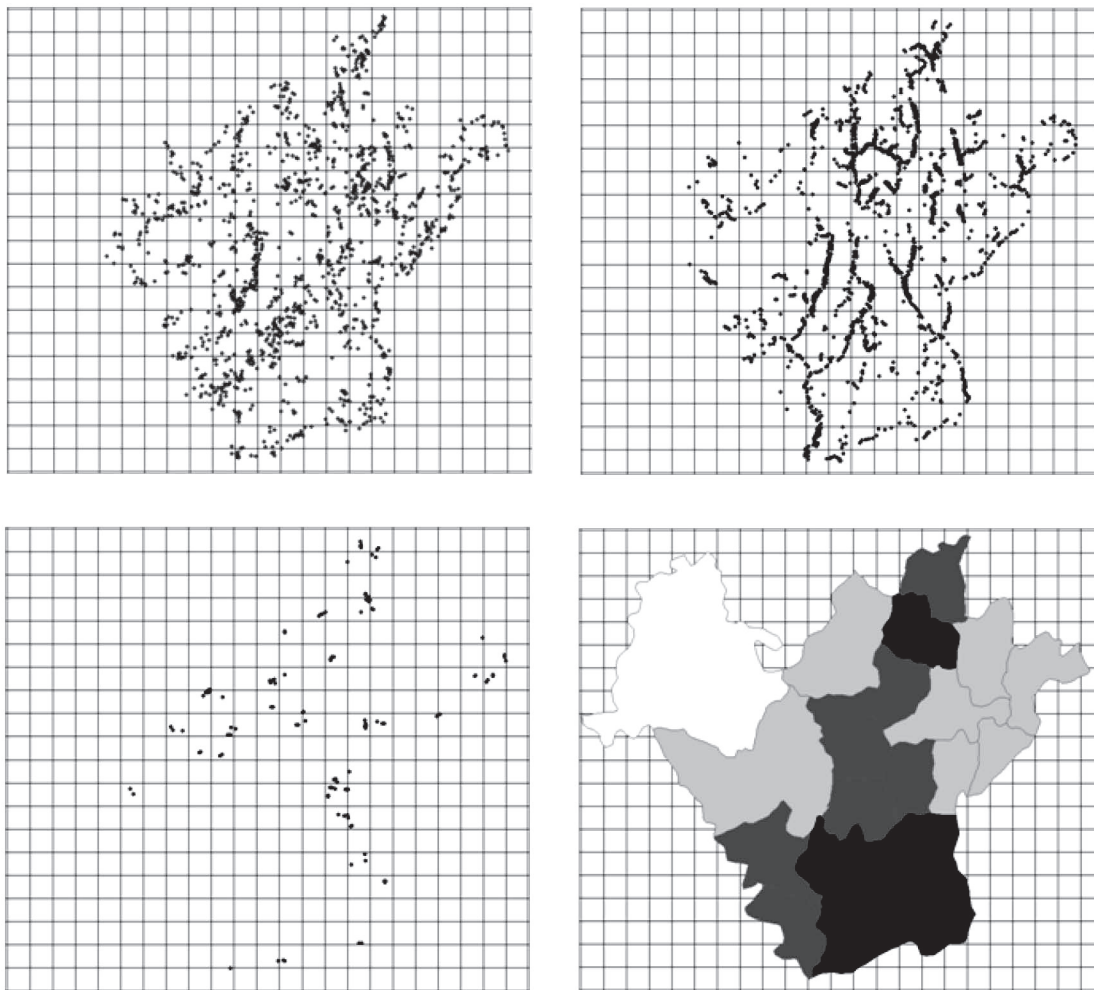


图4 攀西地区滑坡、泥石流、崩塌分布以及危险度区划数据

(左上图为滑坡分布，右上图为泥石流分布，左下图为崩塌分布，右下图为危险度区划)

GL154050190220301, GL154050190220401。

通过地理格网编码能够实现滑坡分布、泥石流分布、崩塌分布以及危险度区划数据的空间集成,便于在统一地理格网上直观展示地质灾害数据空间分布的细节信息,易于用户更有针对性地在地理格网尺度上的数据检索、数据获取、数据分析和制图输出等应用。

4.3 编码的作用分析

基于国家地理格网的陆地表层数据编码的基本作用有以下几个方面。

(1)数据资源管理。属性码段的设计考虑到了数据的加工程度、时间范围、空间范围、分辨率精度、生产方式信息,通过上述信息便于确定数据服务等级与掌握数据资源本底状况。

(2)数据检索访问。通过地理格网进行空间索引,结合数据库等开发工具,可以形成一套格网化数据管理系统,便于数据的空间定位、精细查询,实现高效的数据共享服务。

(3)数据资源集成。以往历史数据多以市(县)级行政区域或自然地理单元(如黄土高原、长江三角洲地区等)积累。无论采用行政区划单元还是自然地理单元对历史数据集成都会存在较多问题,而地理格网通过多级格网系统可以实现对历史数据的无缝集成。

(4)数据资源格网化处理。社会经济等统计数据一般是非空间化、非结构化的数据。在数据集成管理时,往往需要将过去的社会经济等统计数据可进行格网化处理(空间化),这样既便于在空间上浏览和分析社会经济数据的特征,又便于社会经济数据与地理空间数据的集成管理。陆地表层数据的格网化数据编码可为此类数据处理的结果数据集进行编码描述。

(5)数据资源空间分析。经过格网化编码的地理空间数据,能够在统一的、符合我国国情的空间参考下进行多种空间分析,例如空间叠加分析、缓冲区分析、距离优化分析等。结合相关专业数据处理模型,可开展更深入的模型运算分析,例如水文分析、地貌分析、侵蚀分析、土地覆被变化分析、城镇化扩张分析等。

(6)标识引用。经过格网化编码的数据将具有唯一的编码信息。这将有助于数据资源的标识和引用。例如,可将数据资源进行数字对象标识(DOI)注册,并允许使用该数据资源的人对该编码进行引用。同时,在数据标识和引用的基础上,可以建立通过DOI关联的数据资源与相关期刊学术论文的关联,为未来的知识集成提供支持。

5 结语

本文面向科学数据的集成共享,提出一种地理空间数据格网化编码的应用方案。该方案从我国国情出发,借鉴了《国家地理格网标准GB/T 12409-2009》中格网编码方法,建立与国家科技基础条件平台——地球系统科学数据共享平台数据分类之间的连接关系,形成适合于陆地表层数据的格网化编码方案。分别以中科院生态系统观测网络中的盐亭站点状数据和攀西地区地质灾害空间面状数据为对象,给出了相应的格网化数据编码实例。初步分析认为,该格网化编码能够在科学数据资源管理、数据检索访问、数据资源集成、数据资源格网化处理、数据资源空间分析、数据标识与引用等方面发挥作用。

参考文献

- [1] 周成虎,欧阳,马廷.地理格网模型研究进展[J].地理科学进展,2009,28(5):657-662.
- [2] 地理格网GB/T 12409-2009[S].中华人民共和国国家标准,2009.
- [3] 曾梦歧,谭平嶂,陈剑锋.美军GIG3.0 进展研究[J].信息安全与通信保密,2011(12): 50-53.
- [4] 国际农业研究咨询顾问集团空间信息协会[EB/OL]. [2014-06-30]. <http://srtm.csi.cgiar.org/SELECTION/inputCoord.asp>.
- [5] 廖顺宝,孙九林.基于GIS的青藏高原人口统计数据空间化[J].地理学报,2003,58(1):25-33.
- [6] 张超,万庆,张继权,等.基于格网数据的洪水灾害风险评估方法——以日本新川洪灾为例[J].地球信息科学,2003(4):69-73.
- [7] 丁志雄,李纪人,李琳.基于GIS格网模型的洪水淹没分析方法[J].水利学报,2004(6):56-60.

(下转第93页)

- [17] Wu Xianchao, Matsuzaki Takuya, Tsujii Jun'ichi. SMT Systems in the University of Tokyo for NTCIR-9 PatentMT [C] //Proceedings of NTCIR-9 Workshop Meeting, Tokyo: National Institute of Informatics, 2011: 666-672.
- [18] Zheng Zhongguang, Ge Naisheng, Meng Yao, et al. HPB SMT of FRDC Assisted by Paraphrasing for the NTCIR-9 PatentMT [C] //Proceedings of NTCIR-9 Workshop Meeting, Tokyo, National Institute of Informatics, 2011: 679-683.
- [19] Xiao Tong, Li Qiang, Lu Qi, et al. The NiuTrans Machine Translation System for NTCIR-9 PatentMT [C] //Proceedings of NTCIR-9 Workshop Meeting, Tokyo: National Institute of Informatics, 2011: 593-599.
- [20] Xiong Hao, Song Linfeng, Meng Fandong, et al. The ICT's Patent MT System Description for NTCIR-9 [C] //Proceedings of NTCIR-9 Workshop Meeting, Tokyo: National Institute of Informatics, 2011: 673-678.
- [21] He Yanqing, Shi Chongde, Wang Huilin. ISTIC Statistical Machine Translation System for Patent Machine Translation in NTCIR-9 [C] //Proceedings of NTCIR-9 Workshop Meeting, Tokyo: National Institute of Informatics, 2011: 634-637.
- [22] Peter F Brown, Jennifer C Lai, Robert L Mercer. Aligning Sentences in Parallel Corpora [C] //Proceedings of 29th Annual Meeting of the ACL, Berkeley: Association for Computational Linguistics, 1991: 169-176.
- [23] William A Gale, Kenneth W Church. A Program for Aligning Sentences in Bilingual Corpora [C] //Proceedings of 29th Annual Meeting of the ACL, Berkeley: Association for Computational Linguistics, 1991: 177-184.
- [24] Martin Kay, Martin Röchdisen. Text-translation Alignment [J]. Computational Linguistics, MIT Press Cambridge, MA, USA, 1993, 19(1): 121-142.
- [25] Stanley F Chen. Aligning Sentences in Bilingual Corpora Using Lexical Information [C] // Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics: Association for Computational Linguistics, 1993: 9-16.
- [26] Michel Simard, Pierre Plamondon. Bilingual Sentence Alignment: Balancing Robustness and Accuracy [J]. Machine translation, Kluwer Academic Publishers, 1998, 13(1): 59-80.
- [27] 刘昕, 周明, 朱胜火, 等. 基于自动抽取词汇信息的双句子对齐 [J]. 计算机学报, 1998, 21(8): 151-158.
- [28] Robert C Moore. Fast and Accurate Sentence Alignment of Bilingual Corpora [J]. Machine translation: From Research to Real User, 2002, 2499: 135-144.
- [29] Kishore Papineni, Salim Roukos, Todd Ward, et al. BLEU: A Method for Automatic Evaluation of Machine Translation [C] // Proceedings of the 40th Annual Meeting of the ACL, Philadelphia: Association for Computational Linguistics, 2002: 311-318.
- [30] Andreas Stolcke. SRILM-An Extensible Language Modeling Toolkit [C] // Proceedings of International Conference on Spoken Language Processing, Denver: International Speech Communication Association, 2002: 901-904.

(上接第78页)

- [8] 刘明亮, 唐先明, 刘纪远. 基于1 km格网的空间数据尺度效应研究 [J]. 遥感学报, 2001, 5(3): 183-189.
- [9] 国家地球系统科学数据共享平台. 地球系统科学数据共享数据分类标准 [S]. 2013.
- [10] 徐冠华. 实施科学数据共享, 增强科技竞争力 [J]. 中国基础科学, 2003(1): 5-9.
- [11] 廖顺宝, 蒋林. 地球系统科学数据分类体系研究 [J]. 地理科学进展, 2005, 24(6): 93-98.
- [12] 王卷乐, 林海, 冉盈盈, 等. 面向数据共享的地球系统科学数据分类探讨 [J]. 地球科学进展, 2014, 29(2): 265-274. DOI: 10.11867/j.issn.1001-8166.2014.02.0265.
- [13] 高美荣. 盐亭站1985-1994年旱地水分、养分动态观测数据集 [EB/OL]. [2014-06-30]. <http://www.geo-data.cn/Portal/metadata/viewMetadata.jsp?id=100101-610041.10067>.