

文章编号: 1007-4619(2008)01-0009-06

基于语义的地理信息分类体系对比分析

张雪英, 闫国年

(南京师范大学 地理信息系统江苏省重点实验室, 江苏 南京 210024)

摘 要: 地理信息分类体系之间的语义不一致性, 被认为是影响地理信息系统实现在语义层次上信息共享的最大障碍。首先, 本文在阐述中国地理信息分类体系之间参照模式的基础上, 提出了一种基于语义的地理信息分类体系对比分析方法。其次, 对中国现行的几种地理信息标准分类体系进行了详细的对比分析。最后, 为中国地理信息分类体系的编制和修订提出了一些建议。

关键词: 地理信息分类体系; 语义分析; 地理信息系统

中图分类号: P208 **文献标识码:** A

1 引 言

地理信息是指用数字、文字、符号、语言等介质来表征的关于地球表层自然资源和环境要素与人文要素相互作用与关系的内容、数量或特征。地理信息必须按照一定的规律进行分类与编码, 将其有序地存入计算机, 才能进行存储、管理、检索、输出与交换等操作。地理信息分类体系是地理数据标准化建设, 地理信息系统开发, 以及地理信息系统数据共享与互操作的基础^[1]。20 世纪 80 年代之后, 中国地理信息系统的应用需求急剧增长, 各种地理信息分类体系也随之应运而生。中国现行的地理信息分类体系可以划分为三个层次: (1) 国家标准。例如, 《国土基础信息数据分类与代码》(GB/T 13923-92)^[2]、《1:500 1:1000 1:2000 地形图要素分类与代码》(GB 14804-1993)^[3] 和《1:5000 1:10000 1:25000 1:50000 1:100000 地形图要素分类与代码》(GB/T 15660-1995)^[4]; (2) 行业标准。例如, 《城市基础地理信息系统 1:500 1:1000 1:2000 地形图要素分类与代码》(CJJ 100-2004)^[5]; (3) 专用分类体系, 针对某一个或者某些特定的地理信息系统而制定的地理信息分类编码。专用分类体系数量众多, 本文不再一一列举。本文将国家标准和行业标准统称为“标准分类体系”。中国各种地理信息分类体系通常采用“由

下往上”的参照模式(见图 1), 使分类体系之间以最自然的、最简单的方式达到一定程度的语义一致性, 从而使地理信息系统具备一定的语义信息共享能力。但是, 由于考虑问题的出发点, 或者对地理客观现象的描述方式不同, 中国地理信息标准分类体系之间的语义不一致性现象非常普遍。

从理论上讲, “标准化”是彻底解决分类体系语义不一致问题的最简单途径。但事实上, 没有一种分类体系能够适用于任何地理信息系统。即使存在这样的分类体系, 那么它与现行地理信息分类体系之间仍然存在语义不一致的问题。建立分类体系的对照关系是目前国内外普遍采用的解决办法。两个分类体系的对照关系包括所有来自两个分类体系的具有相同或者相近语义的类目之间的转换关系^[6]。因此, 借助分类体系的对照关系, 可以帮助从语义的角度对地理信息分类体系进行对比分析。

本文试图探讨以分类体系的对照关系为基础的地理信息分类体系对比分析方法, 并将之应用于中国地理信息标准分类体系的比较分析。最后, 从提高地理信息分类体系的语义一致性出发, 提出一些建议, 以供中国地理信息分类体系编制和修订时参考。

2 对比分析方法

人工建立分类体系的对照关系, 对时间、人力、

收稿日期: 2006-08-23 修订日期: 2006-11-02

基金项目: 南京师范大学重点科研基金资助项目(编号: 2006105XGQ0051)。

作者简介: 张雪英(1970—)女, 副教授, 博士, 毕业于南京理工大学计算机应用专业, 现从事空间数据挖掘和地理信息系统等方面的研究, 发表学术论文 18 篇, 出版专著 1 部。Email: zhangsnow@163.com

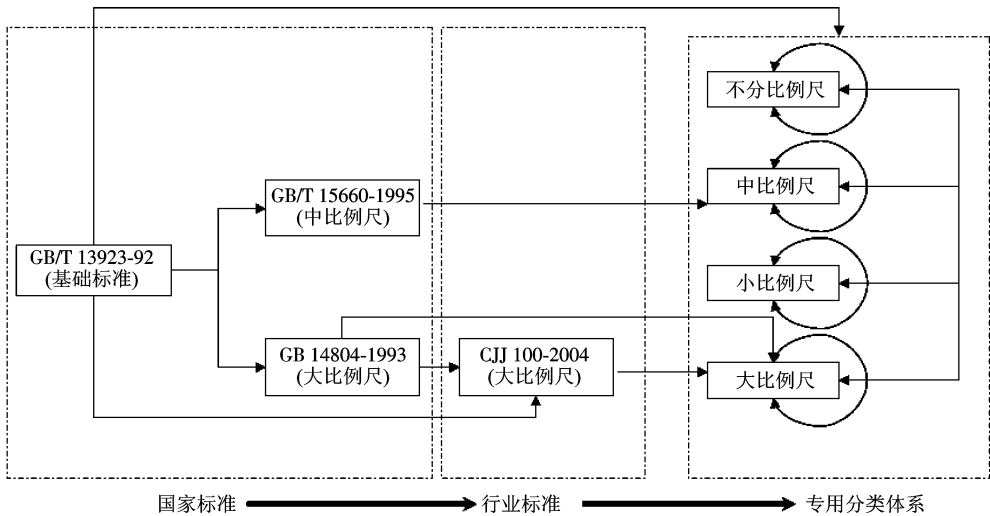


图 1 中国地理信息分类体系的参照模式

Fig 1 Reference model of Chinese geographical classification schemes

物力和财力都有较高的要求。张雪英和闰国年提出了一种应用字面相似度计算方法和语义对照模式,自动生成地理信息分类体系对照关系的方法。实验表明,该方法不仅灵活方便、效率高,而且能够达到较高的准确性^[7]。本文将采用根据此方法开发的地理信息分类体系自动转换系统 (SDCC1.0) 结合人工辅助判断,建立各种地理信息标准分类体系之间的对照关系。表 1 为地理信息分类体系的对照关系样例。其中“等同”表示两个类目的语义完全相

同,而“相关”表示两个类目语义非常相近但不相同。“类名是否相同”用于标识两个类目的类名在字面上是否一样。由于分类体系的类名通常省略与上位类类名相同的部分,如果两个类名不相同,还必须进一步判断它们与上位类类名组合之后的情况。例如,表 1 中的最后一条对照关系中,“收费站”和“高速公路收费站”被认为在字面上是相同的。

分类代码和类名是定义和表达分类体系中各个类目语义内容的主要成分,而对照关系则表示不同

表 1 地理信息分类体系对照关系样例

Table 1 Mapping results of two geographical classification schemes

GB 14804-1993		GB/T 13923-1992		语义关系	类名是否相同
代 码	类 名	代 码	类 名		
1	测量控制点	1	测量控制点	等同	是
11	平面控制点	11000	平面控制点	等同	是
111	三角点	11020	三角点	等同	是
1111	一等	11031	一等	等同	是
1112	二等	11032	二等	等同	是
1113	三等	11033	三等	等同	是
1114	四等	11034	四等	等同	是
115	导线点	11030	导线点	等同	是
1151	一级	11031	一等	等同	否
1152	二级	11032	二等	等同	否
1153	三级	11033	三等	等同	否
117	埋石图根点	13052	埋石图根点	等同	是
118	不埋石图根点	13052	埋石图根点	相关	否
12	高程控制点	12000	高程控制点	等同	是
431	高速公路	42010	高速公路	等同	是
4311	收费站	95250	高速公路收费站	等同	是

分类体系的类目之间的语义关系。因此,从语义的角度进行分类体系的对比分析,可以采用定性分析和定量分析相结合的方法。

(1) 定性分析。主要考察两个分类体系的编码方式。如果两个分类体系采用相同的编码方式,或者一方是对另一方编码的扩展,就有可能获得较好的语义一致性。

(2) 定量分析。通常情况下,分类体系的定量对比分析是非常困难的。到目前为止,人们尚未提出一些比较有效的评价指标。本文在借鉴信息检索领域最常用的检索性能评价指标 $F_1^{[8]}$ 的基础上,定义了一系列定量评价指标。

假设 A 和 B 分别表示两个地理信息分类体系, I_A 表示 A 的类目总数, I_B 表示 B 的类目总数。则评价指标的定义如下:

转换系数 (S) 是指两个分类体系中语义相同或者相近的类目所占比重,用于从整体上反映两个分类体系类目设置的相似性。

$$S = \frac{2 \times \text{对照关系的数量}}{I_A + I_B} \tag{1}$$

按照类目之间语义的相关程度,转换系数可以分为等同系数和相关系数。等同系数 (E) 是指两个分类体系中语义相同的类目所占比重;相关系数 (R) 是指两个分类体系中语义相近的类目所占比重。等同系数越高,说明两个分类体系重复的类目越多;相关系数越高,说明两个分类相关的类目较多。具体定义如下:

$$E = \frac{2 \times \text{等同对照关系的数量}}{I_A + I_B} \tag{2}$$

$$R = \frac{2 \times \text{相关对照关系的数量}}{I_A + I_B} \tag{3}$$

按照类名的字面相似程度,转换系数又可以分

为字面相同系数和字面不相同系数。字面相同系数 (L) 是指两个分类体系中语义相同且类名字面相同的类目所占比重;字面不相同系数 (N) 是指两个分类体系中语义相同或者相近但类名字面不相同的类目所占比重。字面不相同系数反映了两个分类体系类目描述的语义不一致性程度。具体定义如下:

$$L = \frac{2 \times \text{字面相同对照关系的数量}}{I_A + I_B} \tag{4}$$

$$N = \frac{2 \times \text{字面不相同对照关系的数量}}{I_A + I_B} \tag{5}$$

3 地理信息标准分类体系对比分析

3.1 定性分析

《国土基础信息数据分类与代码》(GB/T13923-1992) 是基础标准,适用于各种比例尺、各种专业领域、不同内容的空间定位基础地理信息系统^[9]。GB/T13923-1992 共分 9 大类、4 个分类级别,分类代码由 5 位数字码组成。GB14804-1993 用于规定 1:500 1:1000 1:2000 地形图要素分类与代码,共分 9 大类、4 个分类级别,分类代码由 4 位数字码组成。CJJ100-2004 适用于城市基础地理信息系统中城市空间基础数据的获取、加工、建库、更新和系统建设、管理、维护及数据分发服务等工作。该规范制定的 1:500 1:1000 1:2000 地形要素分类与代码,共分 9 大类、5 个类别级别,分类代码的前 4 位采用 GB14804-1993 中的代码,扩展的第五位是细分码,第六位是辅助码。GB/T15660-1995 用于规定 1:5000 1:10000 1:25000 1:50000 1:100000 地形图要素分类与代码,共分 9 大类、4 个分类级别,分类代码采用 4 位数字码组成。表 2 总结了我国地理信息标准分类体系在分类编码方式上的差异。

表 2 中国地理信息标准分类体系编码方式的差异
Table 2 Encoding differences of standard geographical classification schemes

标准	编码扩展	大类数目	大类次序	分类级别	分类代码	类目总数
GB/T13923-1992	—	9	—	4	5 位数字码	661
GB14804-1993	无	9	与 GB/T13923-1992 不相同	4	4 位数字码	663
CJJ100-2004	GB14804-1993	9	与 GB/T13923-1992 不相同与 GB14804-1993 相同	5	6 位数字码	676
GB/T15660-1995	无	9	与 GB/T13923-1992 不相同	4	4 位数字码	603

从大类设置上看,中国各种地理信息标准分类体系的大类数目完全相同,而大类次序存在一些差异。GB/T13923-1992 的大类设置并没有被其他标准分类体系所采用,而这些分类体系的大类设置却大同小

异。从分类编码来看,各种分类体系虽然都采用数字码(位数不同),但是分类编码之间并不存在直接转换关系。因此,仅仅根据分类代码来分析两个分类体系的类目之间的语义关系,显然具有相当大的难度。

从类目设置规模来看, 各个分类体系的差异较小。

3 2 定量分析

3 2 1 基础标准与其他标准分类体系之间的对比分析

GB/T13923-1992是中国颁布时间最早的标准

地理信息分类体系, 是适用于各种比例尺的基础标准。因此, 将 GB/T13923-1992 与其他标准分类体系进行对比分析具有非常重要的意义。表 3 描述了 GB/T13923-1992 与其他标准分类体系的定量分析结果。

表 3 GB/T13923-1992 与其他标准分类体系的定量对比分析结果
Table 3 Results of quantitative comparison between GB/T 13923-1992 and other standard schemes

	GB140804-1993与 GB/T13923-1992	GB15660-1995与 GB/T13923-1992	CJ1100-2004与 GB/T13923-1992
对照关系数	484	491	477
其中: 等同关系	255	308	113
相关关系	229	183	364
字面相同	213	243	78
字面不相同	271	248	399
转换系数 (C)	73.3%	77.7%	71.4%
其中: 等同系数 (E)	38.5%	48.7%	16.9%
相关系数 (R)	34.6%	30.0%	54.5%
字面相同系数 (L)	32.2%	38.5%	11.7%
字面不相同系数 (N)	40.9%	39.2%	59.7%

从表 3 可以看出, GB140804-1993、GB15660-1995、CJ1100-2004 都是针对特定比例尺的地理信息分类体系, 与 GB/T13923-1992 的转换系数比较接近。GB140804-1993 与 GB/T13923-1992 的等同系数和字面相同系数差异不大, 说明两个分类体系的语义一致性良好。GB15660-1995 与 GB/T13923-1992 的等同系数高达 48.7%, 但是字面相同系数却比较低。因此, 可以判断两个分类体系语义不一致现象比较严重。CJ1100-2004 和 GB/T13923-1992 的等同系数和相关系数分别为 16.9% 和 54.5%, 说明两个分类体系虽然相关, 但是重复的类目却比较少。GB140804-1993、GB15660-1995、CJ1100-2004 与 GB/T13923-1992 的相

关系数都比较高, 主要原因在于 GB/T13923-1992 是不分比例尺的基础标准, 而其他 3 个标准都是针对特定比例尺的分类体系, 其类目设置相对比较细化、专指。

3 2 2 相同和不相同比例尺标准分类体系之间的对比分析

CJ1100-2004 和 GB14804-1993 都是制定 1:500、1:1000、1:2000 地形要素的分类与代码。GB14804-1993 是国家标准, CJ1100-2004 是行业标准。GB/T15660-1995 与 GB14804-1993 都是国家标准, 但是分别规定不同比例尺地形图要素的分类与代码。表 4 给出了 CJ1100-2004 和 GB14804-1993、GB/T15660-1995 与 GB14804-1993 之间的定量分析结果。

表 4 相同和不相同比例尺标准分类体系之间的定量对比分析结果
Table 4 Results of quantitative comparison between common scale and different scale standard schemes

	CJ1100-2004与 GB14804-1993	GB/T15660-1995与 GB14804-1993
对照关系数	646	440
其中: 等同关系	339	295
相关关系	307	145
字面相同	223	219
字面不相同	423	221
转换系数 (C)	96.5%	69.5%
其中: 等同系数 (E)	50.6%	46.6%
相关系数 (R)	45.9%	22.9%
字面相同系数 (L)	33.3%	34.6%
字面不相同系数 (N)	63.2%	34.9%

CJJ100-2004 与 GB14804-1993 的转换系数为 96.5%, 且等同系数超过 50%。这说明两个分类体系中, 大约一半的类目是完全相同的, 而另一半类目也有较高的相关性。但是, 字面相同系数却显示两个分类体系的语义不一致现象非常突出。因此, CJJ100-2004 和 GB14804-1993 作为针对相同比例尺的地理信息标准分类体系, 类目设置差异小而语义不一致性现象严重的情况, 对于地理信息系统实现语义信息共享会有较大影响。

从表 4 中数据可以看出, GB/T 15660-1995 与 GB14804-1993 虽然都是针对不同比例尺的地理信息分类体系, 但是相同类目所占比重却比较大, 相关类目比较少。这一现象说明, 不同比例尺的分类体系同样存在较多的相同类目, 完全可以通过改进达到较好的语义一致性。

4 几点建议

克服地理信息标准分类体系的语义不一致性, 是地理信息系统实现语义信息共享的根本途径。本文根据前文中中国地理信息标准分类体系之间的对比分析结果, 特提出以下几点建议。

(1) 修订 GB/T 13923-1992 的大类设置, 使之尽可能与其他各项标准分类体系的大类一致, 包括分类代码和类名。从理论上讲, 修订基础标准会带来比较大的影响。但是, 由于其他标准分类体系都采用相同的大类设置, 在这种情况下, 修订基础标准反而比较合适。

(2) CJJ100-2004 与 GB140801-1993 类目设置差异比较小, 通过修订可以达到较好的语义一致性。最简单的做法就是将两个分类体系中语义相同的类目尽可能用相同的类名进行描述。另外, 也可以考虑在完善 GB140801-1993 的情况下, 不再保留 CJJ100-2004。但是, 如果 CJJ100-2004 的应用面较广, 取消 CJJ100-2004 势必带来较大的影响。

(3) 建立地理信息本体, 规范相关术语的定义及表达形式, 并将之应用于各种标准分类体系。类目描述方式的不一致是导致分类体系语义不一致的主要原因之一, 规范地理信息概念及其表达方式, 有助于提高分类体系之间的语义一致性。

(4) 由权威部门制定地理信息标准分类体系之间的对照关系。修订标准分类体系周期比较长, 而且同时要修订多个标准分类体系会有较大的难度。如果权威部门能够提供统一的地理信息标准分类体

系之间的对照关系, 那么地理信息系统不仅可以降低实现语义信息共享的代价, 而且可以向用户提供一致的地理信息服务。

(5) 尽可能参考国际标准化组织 (ISO) 地理信息技术委员会 (ISO/TC211) 中有关标准, 特别是地理信息编码 (ISO/DIS 19118) 中基于 XML 的编码规则和编码转换规则, 编制和修订中国地理信息分类体系, 以便于实现具有较高语义一致性的地理数据交换与共享。

参考文献 (References)

- [1] Li X T, He J B, Bi J T. Application Schema of National Resources and Environment Information Classification and Codes [J]. Acta Geographica Sinica, 2002, 57 (Supplement): 9-17. [李新通, 何建帮, 毕建涛. 国家资源环境数据库信息分类编码及应用模式 [J]. 地理学报, 2002, 57 (增刊): 9-17.]
- [2] China State Bureau of Technical Supervision. Classification and Codes for the National Land Information National Standard of the People's Republic of China (GB/T 13923-1992) [S]. 1992 [国家技术监督局. 国土基础信息数据分类与代码 (GB/T 13923-1992) [S]. 1992.]
- [3] China State Bureau of Technical Supervision. Classification and Codes for the Features 1:500 1:1000 1:2000 Topographic Maps [S]. National Standard of the People's Republic of China (GB/T 14804-1993). [国家技术监督局. 1:500 1:1000 1:2000 地形图要素分类与代码 (GB14804-1993) [S]. 1993.]
- [4] China State Bureau of Technical Supervision. Classification and Codes for the Features 1:50000 1:100000 1:250000 1:500000 1:1000000 Topographic Maps [S]. National Standard of the People's Republic of China (GB/T 15660-1995). 1995. [国家技术监督局. 1:50000 1:100000 1:250000 1:500000 1:1000000 地形图要素分类与代码 (GB/T 15660-1995). 1995.]
- [5] Beijing Institute of Surveying and Mapping et al. Technical Specification for Urban Fundamental Geographic Information System (CJJ100-2004) [S]. Beijing: China Architecture & Building Press, 2004. [北京市测绘设计院等. 城市基础地理信息系统技术规范 (CJJ100-2004) [S]. 北京: 中国建筑工业出版社, 2004.]
- [6] Zhang X Y. Concept Integration of Document Databases Using Different Indexing Languages [J]. Information Processing & Management, 2006, 42: 121-135.
- [7] Zhang X Y, Lv G N. Approach to Automatic Conversion of Geographic Information Classification Schemes [R]. Jiangsu Provincial Key Lab of Geography Information Science, Nanjing Normal University, 2007. [张雪英, 闻国年. 地理信息分类体系自动转换方法研究 [R]. 南京师范大学江苏省地理信息科学重点实验室, 2007.]
- [8] van Rijsbergen C J. Information Retrieval. London: Butterworths, 1979.
- [9] Liu R M, Jiang J T. Classification Principles and Methods of

Geographic Information: A Case Study on Basic Geographic
Information[J]. Science of Surveying and Mapping, 2004

29(7): 84—87. [刘若梅, 蒋景瞳. 地理信息的分类原则与方
法研究[J]. 测绘科学, 2004, 29(7): 84—87.]

Comparison Analysis on Semantic based Geographic Classification Schemes

ZHANG Xue-ying, LU Guo-nian

(Jiangsu Provincial Key Lab of Geography Information Science Nanjing Normal University, Jiangsu Nanjing 210024, China)

Abstract: Semantic inconsistency between geographical classification schemes is believed as the most serious problem for heterogeneous geographical information systems to implement semantic information sharing. Since 1980's, four standard Chinese geographical classification schemes have been constructed, i.e., Classification and Codes for the National Land Information (GB/T13923-1992), Classification and Codes for the Features 1:500 1:1000 1:2000 Topographic Maps (GB/T14804-1993), Classification and Codes for the Features 1:5000 1:10000 1:25000 1:50000 1:100000 Topographic Maps (GB/T15660-1995), and Classification and Codes for the Features 1:500 1:1000 1:2000 Topographic Maps for Urban Fundamental Geographical Information System (CJJ100-2004). Individual special geographical classification schemes are normally constructed for specific geographical information systems. GB/T13923-1992 is a general and basic standard geographical classification scheme. All other geographical classification schemes are suggested to be compatible with GB/T13923-1992.

Geographical information systems are usually inconsistent in terms of their content and geographical classification schemes. A single geographical concept might be represented in different ways so that geographical information systems cannot initially implement semantic information sharing. Semantic mapping is considered as a tool to state semantic relations between entities belonging to different classification schemes. It has been proven to most effectively resolve semantic heterogeneity between geographical classification schemes. Although intellectual methods can attain good effects, there are still problems if there are insufficient time, money and experts.

This paper firstly describes a method for identification of entities from different geographical classification schemes. Based on semantic relations, parameters for the qualitative and quantitative analysis of their semantic consistency are defined. And then the basic architectures and coding measures of standard geographical classification schemes are compared and analyzed. The results of the quantitative comparison between GB/T13923-1992 and other standard schemes show that GB/14803-1993 and GB/T13923-1992 have the best semantic consistency. The results of the quantitative comparison between common scale and different scale standard schemes show that GB/14803-1993 and CJJ100-2004 have the worst semantic consistency, but GB/T15660-1995 and GB/14803-1993 can achieve better semantic consistency. Finally, some suggestions are given for construction and modification of Chinese standard geographical classification schemes. For example, standard schemes should define identical main classes and item names; authorized semantic mapping of standard geographical information schemes should be published, and ontologies may be able to solve the problem of semantic inconsistency between photographic geographical information schemes naturally and economically.

Key words: geographical classification scheme; semantic analysis; geographical information system