资源数据文化内涵与旅游价值数据挖掘规范

数字文化旅游平台规范建设课题组

二〇一六年六月

目　　录

1 范围

本规范规定了文化旅游资源数据的描述要求。

本规范适用于“数字文化旅游共性支撑技术研发与区域资源集成应用示范”课题应用，其他相关领城也可参考使用。

在不同的实际旅游目的地进行数据采集工作时，所采集的数据内容根据具体情况和需求可以少于或多于本规范内容，但应达到相关技术要求。

2 规范性引用文件

下列文件中的条款通过本规范的引用而成为本规范的条款。凡是注明日期的引用文件，其随后所有的修改版（不包括勘误的内容）或修订版均不适用于本规范。凡是不注明日期的引用文件，其最新版本适用于本规范。

GB/T 17775-2003 旅游区(点)质量等级的划分与评定

GBT 18973-2003 旅游厕所质量等级的划分与评定

LBT 011-2011 旅游景区游客中心设置与服务规范

GBT 26363-2010 民族民俗文化旅游示范区认定

GBT 3358.1-2009 统计学词汇及符号 第1部分：一般统计术语与用于概率的术语

GB/T 7408-2005 数据元和交换格式 信息交换 日期和时间表示法 （idt ISO 8601:2000）

GB/T 7408-1994 数据元和交换格式 信息交换：日期和时间表示法

GB/T 18391.3-2009 数据元的规范与标准化 (idt ISO/IEC 11179.3-2003)

文化旅游资源兴趣点及道路采集规范

3 术语和定义

下列术语和定义适用于本规范。

3.1 旅游价值数据挖掘

旅游价值数据挖掘就是从大量的文化旅游资源类数据中提取潜在有旅游价值的信息，为管理者开发文化旅游产品提供原来不知道的有价值信息。

3.2 旅游价值分类分析

旅游价值分类分析是找出文化旅游资源数据中数据对象的文化共同特点和旅游价值共同点，并按照分类模式将其划分为不同的类，其目的是通过分类模型，将文化旅游资源数据中的数据项映射到某个给定的类别。它可以应用到旅游管理的分类、旅游者的属性和特征分析、旅游者满意度分析、旅游者的出游趋势预测等。

3.3 Web 页挖掘

随着 Internet 的迅速发展及 Web 的全球普及， 使得 Web 上的信息量无比丰富，通过对 Web 的挖掘，可以利用 Web 的海量数据进行分析，收集政治、经济、政策、科技、金融、各种市场、竞争对手、供求信息、客户等有关的信息，集中精力分析和处理那些对企业有重大或潜在重大影响的外部环境信息和内部经营信息，并根据分析结果找出企业管理过程中出现的各种问题和可能引起危机的先兆，对这些信息进行分析和处理，以便识别、分析、评价和管理危机。

3.4 回归分析

回归分析方法反映的是事务数据库中属性值在时间上的特征，产生一个将数据项映射到一个实值预测变量的函数，发现变量或属性间的依赖关系，其主要研究问题包括数据序列的趋势特征、数据序列的预测以及数据间的相关关系等。它可以应用到市场营销的各个方面，如客户寻求、保持和预防客户流失活动、产品生命周期分析、销售趋势预测及有针对性的促销活动等。

3.5 聚类分析

聚类分析是把一组数据按照相似性和差异性分为几个类别，其目的是使得属于同一类别的数据间的相似性尽可能大，不同类别中的数据间的相似性尽可能小。它可以应用到客户群体的分类、客户背景分析、客户购买趋势预测、市场的细分等。

3.6 关联规则分析

关联规则是描述数据库中数据项之间所存在的关系的规则，即根据一个事务中某些项的出现可导出另一些项在同一事务中也出现，即隐藏在数据间的关联或相互关系。在客户关系管理中，通过对企业的客户数据库里的大量数据进行挖掘，可以从大量的记录中发现有趣的关联关系，找出影响市场营销效果的关键因素，为产品定位、定价与定制客户群，客户寻求、细分与保持，市场营销与推销，营销风险评估和诈骗预测等决策支持提供参考依据。

3.7 特征分析

特征分析是从数据库中的一组数据中提取出关于这些数据的特征式，这些特征式表达了该数据集的总体特征。如营销人员通过对客户流失因素的特征提取，可以得到导致客户流失的一系列原因和主要特征，利用这些特征可以有效地预防客户的流失。

3.8 变化和偏差分析

偏差包括很大一类潜在有趣的知识，如分类中的反常实例，模式的例外，观察结果对期望的偏差等，其目的是寻找观察结果与参照量之间有意义的差别。在企业危机管理及其预警中，管理者更感兴趣的是那些意外规则。意外规则的挖掘可以应用到各种异常信息的发现、分析、识别、评价和预警等方面。

4.数据管理

数据管理包括数据目录管理、元数据管理、数据元管理、数据质量管理等模块组成。

4.1 数据目录管理

数据目录定义允许各种类型的用户进行在线定义数据，包括普通用户、部门管理员等，统一由平台管理员审核。数据目录定义可以定义数据的名称、来源、更新周期、责任人、主题、语言、描述等，定义的数据目录可以选择是否可被下载，可以下载的数据会被打包成 Excel 和 XML 两种格式供用户下载使用。另外，数据目录定义时可以选择“是否允许在线浏览”，允许在线浏览的数据可以被用户在线分页查看数据明细，并可以进行在线数据统计、分析、挖掘和数据可视化。

4.2 元数据管理

文化内涵及旅游价值数据挖掘涉及多种信息资源、业务数据库、非结构化数据、业务模型、业务规则，以及信息资源的各种属性特征。因此，必须对这些特征元素进行统一管理，用户可以从技术和业务角度掌握全部信息资源的生产、存储、转换和同步等所有相关活动，提升信息共享，帮助技术人员和业务人员理解每一个数据的来龙去脉。

元数据定义是对数据资源的规范化描述，是按照相关行业的信息标准，从信息资源中抽取出共性特征，组成的一个特征元素集合，是用一组属性描述定义、标识、表达、转换规则和允许值的数据单元，这种规范化描述可以准确和完备的说明信息资源的各项特征，帮助信息资源使用者准确定位资源、准确规范的使用资源。

元数据定义是对数据目录对应的表结构的定义，包括元模型(表)基本信息和元数据(列)信息。元模型信息包括表的名称、中文描述、数据存储方式，其中数据存储方式有两种：大数据中心和远程数据库。元数据信息包括列的名称、数据类型、长度、标签、隐藏设置等。隐藏设置是对数据隐私加密的一种方式，用户可以通过设置元数据的隐藏规则从而实现对数据的加密，比如将元数据为姓名的这一列定义隐藏规则为第二到第三个字符加密，则实际数据中这一列所对应的所有数据的第二到第三个字符将都会被隐藏加密。

4.3 数据源管理

数据源定义可以定义数据源的名称、地址、用户名、密码、连接池信息等，并支持在线测试。数据源按照用途分为私用和公开两种，按照类型分关系型数据库数据源(MySQL、Oracle、SQL Server)和非关系型数据库数据源(HBase)，定义完成的数据源会在数据挖掘平台中自动实例化成最优的数据库连接池，为数据的获取做好准备。

4.4 数据质量管理

数据质量是从数据整合、数据预处理、资源入库、资源监控、资源利用等数据处理流程环节入手，建立完善的数据生命周期管理与数据质量管控机制，是对数据从获取、清洗、转换、关联、存储、使用等生命周期的每个阶段里可能引发的各类数据质量问题，进行识别、度量、监控、预警等一系列质量管理的活动。数

据质量管理是循环管理过程，其终极目标是通过可靠的数据，提升数据在政府、企业决策等业务中的使用价值。

数据质量作为大数据挖掘的核心之一，具有多重属性，其基本质量特性主要包括：完整性、一致性、准确性和及时性等四个方面，要对数据质量进行较好地控制，就必须对数据的四个基本质量特性进行很好了解，从而在各个方面采取措施，杜绝数据质量问题的出现，使数据监控工作能够真正达到控制数据质量的目的。

a.数据完整性检测

数据完整性检测主要包括数据目录对应的数据集是否存在；实际数据集是否存在；数据集中的主键、约束、数据类型相对于元数据是否有缺失等。

b. 数据一致性检测

数据一致性检测指实际数据的值是否和元数据定义的标签的验证规则一致。

c. 数据准确性检测

数据准确性检测主要指数据集中的数据是否有乱码；实际数据的数据类型是否和元数据中定义的相同。

d. 数据及时性检测

数据及时性检测指数据的更新周期是否符合数据目录中定义的更新周期。

e. 质量评分

针对数据集的缺失、元数据的不匹配、数据的不准确等数据质量问题，数据管理平台制定了客观的评分规则，对每个数据目录进行统一打分，并提供了完善的管理系统在线预览数据质量评分排名、有问题的数据、各项数据质量问题的明细等。

5 数据同步

数据同步包括两种同步方式：数据交换和互联网数据采集。

5.1 数据交换

针对数据来源一般是多个业务系统，数据整合难度大，错误数据、不完整数据、重复数据等“脏数据”较多，数据量大，转换任务并发量大等问题，大数据平台要提供完善的分布式数据转换工具与基于 web 的在线任务执行与监控系统，通过建立抽取模型将现有各个系统的数据进行整合、汇总、清洗、转换，并可以实时在线监控交换任务的数据输入量、输出量、成功个数、失败个数、失败原因等。

数据交换支持弹性部署。针对高并发、大数据量的数据转换任务，可以将单个转换任务分配给多个子服务器并行执行，对于转换过程中的每一步操作都有相关的日志记录、转换明细等，并支持事务回滚、数据更新等操作。

5.2 互联网数据采集

针对互联网数据量大、中文分词和向量空间模型导致效率低下等问题，要支持自动化集群管理、任务分

配、负载平衡体系与基于机器学习算法建立的决策树模型的分布式爬虫技术，配合丰富实用的图形化界面，自动生成数据抽取和格式化规则，强力锁定目标数据的内容结构，应对数据结构的变化。通过定向、分类的数据抽取，将相关的数据统一收集到数据管理中心，进行分布式存储。

6. 数据存储与计算

6.1 数据计算

数据计算包括离线计算、内存计算、流式计算等，数据存储包括关系数据库、列式数据库及分布式文件系统等。针对不同的数据类型使用不同的存储方式，保证数据的存储质量。

6.2 离线计算

离线计算一般是批量处理数据库的过程，比如利用 Hadoop 的 MapReduce。

6.3 内存计算

内存计算是将数据是放在在内存中，效率比较高，流式计算为应用提供高效分布式的触发式任务、定时任务等事件驱动程序处理能力。

7. 数据分析与挖掘

7.1 内容

数据分析与挖掘包括但不限于离线分析、在线分析、数据挖掘、搜索引擎、推荐引擎等功能。

在线分析主要面向关系型数据库，离线分析主要指面向 HBase 等非关系型数据库的数据分析。数据分析的主要功能包括对数据的筛选、分组、统计、排序等。

7.2 关系型数据的分析计算

对于关系型数据库，在用户选择分析模型与分析条件后，系统应可以在线快速反馈分析结果。对于用户提交的分析结果，经过管理员审核通过后，系统应能自动生成一套完整的数据目录、元数据与分析视图，并作为一个新的数据集补充到数据挖掘平台内，方便其他用户在线浏览、学习。

7.3 非关系型数据的分析

对于非关系型数据库，分析系统应记录下对应的数据目录、元数据与分析视图，并自动进入审核流程，审核通过的离线分析请求会通过分布式并行计算离线计算分析结果，并能生成一套完整的数据集。

7.4 用户要求

数据挖掘需要用户具有一定的业务基础并且对挖掘算法有个最基本的了解，其分布式并行算法可以满足用户对海量数据的分析与预测。数据挖掘支持用户在线选择数据模型与挖掘算法，并自动建模。平台应能提供多种常用的机器算法可供用户选择使用，比如贝叶斯网络、支持向量机、决策树、隐马尔可夫等，基于这些算法可以建立满足不同业务需求的挖掘模型，实现数据的分析、预测。

7.5 常用分析方法

数据挖掘应能支持分析常用的方法主要有分类、回归分析、聚类、关联规则、特征、变化和偏差分析、

Web 页挖掘等， 便于分别从不同的角度对数据进行挖掘。

7.6 挖掘分析框架

一般数据挖掘为层次结构，分为挖掘操作模块、数据预处理模块、存储控制模块、挖掘库及挖掘库管理模块，数据库和外部文件等。

7.7 设计要求

数据挖掘系统框架的设计应能考虑到一下几点：

(1)数据挖掘系统包括很多方面的操作，这些操作所要求的数据源形式不同、输出不同、所需参数不同，这就使得实现这些操作的各个挖掘操作模块之间必须相对独立。

(2)数据挖掘系统作为一个整体，必须能够协调各个操作模块之间的工作。系统使用挖掘库提供统一的机制来管理各模块所使用的数据源、参数和挖掘结果。

(3)数据挖掘的对象既可能存在于数据库或数据仓库中，也可能存在于文件中，系统应该分别提供处理它们的相应方法。

(4)数据挖掘的结果需要保留。这一方面是因为数据挖掘的目的是支持决策分析;另一方面是为了方便重新挖掘、增量挖掘。

(5)作为一个支持决策分析的系统，其使用者不是计算机工作者，而是决策者，系统应该提供友好的界面。

7.8 数据预处理

数据预处理模块的主要功能是定义数据源、格式化数据源以及过滤数据源。

7.8.1 泛化

泛化就是将相关数据或概念泛化到更高级的层次上。本系统集成的泛化算法是 GDBR。该算法的特点是:对比其他算法(如 LCHR，AOG 等)，它有最好的时间复杂度 O(n)以及很好的空间复杂度 O(c)。

7.8.1 数据清洗

数据清洗的主要工作就是准确、高效地检测出数据库中的相似重复记录。系统使用一种基于 N-Gram[2]的检测相似重复记录的综合方法，能处理常见的拼写错误，如插入、删除、交换、替换和单词的交换等。为了消除基本算法在检测精度上的一些不足，系统采用了经过改进的算法，在实现中运用了统计学原理较好地去除了噪声，并综合应用了正向和逆向重复矩阵，提高了插入/删除错误的检测率。

7.9 数据挖掘模块

不同的挖掘操作模块负责不同的数据挖掘操作。它们彼此之间相对独立，共同之处是都受到挖掘库管理模块的管理，通过存储控制模块获得数据，并把结果写入挖掘库。

数据分析挖掘平台主要包括五类功能：自动预测趋势和行为功能、关联分析功能、聚类分析功能、概念描述功能、偏差检测功能，并对基础数据分析进行展现。

7.9.1 自动预测趋势和行为功能

数据挖掘自动在大型数据库中寻找预测性信息，以往需要进行大量手工分析的问题如今可以迅速直接由数据本身得出结论。例如市场预测问题，数据挖掘使用过去有关市场的数据来寻找未来投资中回报最大的用户，其它可预测的问题包括预报破产以及认定对指定事件最可能作出反应的群体。

广义的关联规则可分为 3 类：强规则、例外规则和随机规则。使用强规则(大部分数据服从的规则)可以帮助预料将来的情况。

7.9.2 关联分析功能

数据关联是数据库中存在的一类重要的可被发现的知识。若两个或多个变量的取值之间存在某种规律性，就称为关联。关联可分为简单关联、时序关联、因果关联。关联分析的目的是找出数据库中隐藏的关联网。有时并不知道数据库中数据的关联函数，即使知道也是不确定的，因此关联分析生成的规则带有可信度。

同关联规则一样，挖掘时序模式的问题也源于由记录组成的事务数据库 D，但时序模式主要是对物品(项)在时间上的关联性加以考虑。Golden-Eye 系统集成的时序模式挖掘算法是 Agrawal 等人提出的 AprioriAll 算法，关于算法的说明在此不作赘述。

分类的基本思想是:根据一些已定义好类别的数据的信息，产生一个可以描述数据类别或对未知类别的数据进行分类的分类器。本系统集成的分类算法最终生成的分类器被称为区间分类器(interval classifier)。该算法的特点在于与采用二叉树的决策树分类器相比，它的准确度较高，决策树的深度也不至于过深。

7.9.3 聚类分析功能

数据库中的记录可被化分为一系列有意义的子集，即聚类。聚类增强了人们对客观现实的认识，是概念描述和偏差分析的先决条件。聚类技术主要包括传统的模式识别方法和数学分类学。80 年代初，Mchalski 提出了概念聚类技术，其要点是，在划分对象时不仅考虑对象之间的距离，还要求划分出的类具有某种内涵描述，从而避免了传统技术的某些片面性。

系统集成的考虑综合因素的聚类方法吸收了一些现有聚类算法的优点，使用层次聚类方法的框架，综合考虑簇之间的距离和簇中对象的密度来决定两个簇是否应该合并。它吸收了 CURE 算法中采用多个代表点来表示簇的方法，因而能够有效地识别特殊形状的簇。为了增强处理大数据量的能力，在使用层次聚类法之前，算法将对象所分布的数据空间划分成数据单元，计算统计信息后得到初始的簇。最后，算法利用索引对数据库中的所有对象进行标记。该算法的主要步骤如下:(1)取样;(2)划分数据单元;(3)消除噪声;(4)利用距离和密度判断、合并簇;(5)识别 outlier;(6)标记数据。

7.9.4 概念描述功能

概念描述就是对某类对象的内涵进行描述，并概括这类对象的有关特征。概念描述分为特征性描述和区

别性描述，前者描述某类对象的共同特征，后者描述不同类对象之间的区别。生成一个类的特征性描述只涉及该类对象中所有对象的共性。

生成区别性描述的方法很多，如决策树方法、遗传算法等。

7.9.5 偏差检测功能

数据库中的数据常有一些异常记录，从数据库中检测这些偏差很有意义。偏差包括很多潜在的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。偏差检测的基本方法是，寻找观测结果与参照值之间有意义的差别。

该功能用到的挖掘方法有：基于统计（statistical-based)的方法、基于距离(distance-based)的方法、基于偏差(deviation-based)的方法、基于密度(density-based)的方法、高维数据的异常探测。

7.10 存储控制模块

系统假设数据源存放在数据库中，由存储控制模块对数据库统一进行操作。对于存放在外部文件中的数据，需要使用数据库管理系统提供的导入工具把数据导入数据库以后再进行挖掘操作。当前，系统的数据源存放在数据库中，从可移植性的角度考虑，我们对存储控制进行封装高于，这是因为数据挖掘应用不同于一般的数据库应用程序，它对数据库的访问频繁，而每次对数据库的访问都会耗费一定的时间和资源。对于数据挖掘操作来说，对大数据量的处理能力和处理效率是一个根本的问题，所以，由系统来进行缓冲和内存索引就非常重要。存储控制模块的功能主要体现在 3 个方面：

1.对连接数据库、管理外部文件以及交换外部文件和内存的内容等较为底层的操作进行封装。

2.负责缓冲管理。具体地说，该模块为数据源、数据挖掘中间结果以及挖掘结果分别申请缓冲区，并保证其驻留在内存中。

3.提供简单的数据格式转换。不同于数据预处理模块提供的数据格式转换，该功能主要弥补关系数据库不能存储不规则格式数据的问题:在向缓冲区中存放数据以前对事务记录进行重新拼接。

7.11 挖掘库及挖掘库管理模块

挖掘库和挖掘库管理是整个系统的核心部分。挖掘库是一个逻辑概念。一个挖掘库存放用户所指定的一系列挖掘操作的所有信息。在系统中，所有的挖掘库都统一存放在数据库中，由系统统一管理。挖掘库所保存的挖掘操作是指包括数据准备和数据挖掘在内的所有操作。在挖掘库中存放的这些操作信息是有顺序的(用户进行这些操作的顺序)。这是因为一个数据挖掘操作在整个知识发现过程中往往不是孤立的，它所使用的数据源常常是另一个数据挖掘操作的结果，而它的挖掘结果又有可能是其他操作的数据源。所以，保留挖掘顺序实际上就是保留了挖掘操作之间的这种关系，这无论对用户理解挖掘结果还是以后重新进行挖掘都是有帮助的。除了操作的名称和顺序以外，挖掘库还保存数据源信息、挖掘操作的参数设置以及挖掘的结果。因此，我们的系统能够很方便地实现把一个挖掘操作的结果作为另一个挖掘操作的输入。我们提供了一套管理挖掘库的操作，这些操作被封装成挖掘库管理模块。图形界面通过调用挖掘库管理模块来完成对挖

掘库的管理。同时，挖掘库管理模块通过调用各个挖掘操作模块来实现挖掘操作。管理挖掘库的所有操作可以被分成以下 4 类。

1.对挖掘库的操作。这组操作主要提供对挖掘库整体的管理。包括连接挖掘库、断开挖掘库、打开挖掘库、增加挖掘库、存储挖掘库、删除挖掘库和查询挖掘库。任何对挖掘库的操作必须在打开了一个挖掘库以后才能进行，而系统的任意运行时刻最多只能打开一个挖掘库。

2.对数据源的操作。这组操作主要用于定义数据源。包括查询数据库信息、增加数据源、查询数据源信息等。

3.对挖掘操作的设置操作。包括增加挖掘操作、查询挖掘操作、设置挖掘操作参数、查询挖掘操作参数等。

4.对挖掘结果的操作。系统实现了对挖掘结果的查询操作。