文章编号:1007-6735(2016)05-0487-10 DOI:10.13255/j.cnki.jusst.2016.05.012

数据显断层与数据隐断层的研究与应用

夏骄雄[1,2,3,4], 韦 琴[2], 缪 慧[3], 吴新林[5], 徐 钊[3], 金 勇[3], 高伟山[2]

(1.上海外国语大学 贤达经济人文学院,上海 200083; 2.上海理工大学 光电信息与计算机工程学院,上海 200093; 3.上海大学 计算机工程与科学学院,上海 200444; 4.上海市教育委员会 信息中心,上海 200003; 5.上海市教育评估院 教育评估研究所,上海 200031)

摘要:建立能够和谐平衡各个信息系统之间数据断层的机制是实现管理决策变革最关键的3大基础之一,也是智能决策支持系统研究领域的重点内容之一.随着互联网+时代的到来,各式各样的数据资源不断积累,数据断层现象在多个领域表现得愈加明显.通过对数据断层理论体系的进一步研究与实践,着重分析微观层面的数据断层现象,一方面用显断层概念描述各系统之间以及系统内部存在的较为明显的断层现象,另一方面用隐断层概念描述各系统之间以及系统内部存在的非明显的断层现象,并在数据显断层中引入缝隙的概念来描述主题无关数据对象,采用数据聚合的技术手段来降低缝隙数据的断层属性,同时在隐断层中引入"有效密度"来形象地描绘数据分布情况,通过数据熔合来减少无效数据占用的空间.最后以上海"动感101"音乐电台的移动客户端应用日志数据为例,分析了电台数据中所存在的数据显断层和数据隐断层现象.

关键词:智能决策支持系统;微观数据断层;数据聚合;数据熔合;缝隙检测;数据显断层;数据隐断层
中图分类号:TP 311.131;G 202 **文献标志码**:A

Research and Application of Dominant Data Faultage and Tacit Data Faultage

XIA Jiaoxiong[1,2,3,4], WEI Qin[2], MIAO Hui[3], WU Xinlin[5], XU Zhao[3], JIN Yong[3], GAO Weishan[2]

(1. *Xianda College of Economics and Humanities, Shanghai International Studies University, Shanghai 200083, China*; 2. *School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*; 3. *School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China*; 4. *Information Centre, Shanghai Municipal Education Commission, Shanghai 200003, China*; 5. *Department of Education Evaluation Research, Shanghai Education Evaluation Institute, Shanghai 200031, China*)

Abstract: As one of the three key bases for the alteration of management decision-making, the establishment and perfection of the mechanism of data faultage between different information systems is also one of the important research content for intelligent decision support systems. With the advent of the Internet +, all kinds of data resources are growing, the data faultage has become more

收稿日期:2016-03-07
基金项目:国家自然科学基金资助项目(40976108,61303097);上海市重点学科建设资助项目(J50103);上海市第二期(2016年)民办高校科研项目(2016-SHNGE-08ZD);上海大学研究生创新基金资助项目(SHUCX070037,SHUCX120105)
第一作者:夏骄雄(1973-),男,研究员.研究方向:数据挖掘、智能决策支持系统、教育信息化、计算机辅助教育.E-mail:jshardrom@shmec.gov.cn

obvious in several areas around our life. For further studies and practices on data faultage intellectual systems, taking the micro data faultage as a main analysis object, the dominant data faultage was defined as the data faultage between different information systems, and the tacit data faultage was defined as the data faultage existing in a single information system. In the dominant data faultage, the concept of pore was introduced as the data object with subject independence, and a data compaction technique was used to reduce the faultage property of pore. In the tacit data faultage, the data density was defined as the distribution of data objects, and a data pressure solution technique was used to reduce the space occupied by data objects. In the paper, the log data of mobile client application of "Shanghai Music Radio FM 101.7" was taken as an example and the dominant data faultage and the tacit data faultage were analyzed finely.

Keywords: *intelligent decision support system*; *micro data faultage*; *data aggregation*; *data fusion solution*; *crack detection*; *dominant data faultage*; *tacit data faultage*

IBM的行业咨询总监赵栩曾经指出:要实现管理决策的变革,关键的3大基础是建设完成符合管理转型和流程优化要求的信息系统,建立完善能够和谐平衡各个信息系统之间数据断层的机制,建构完整满足实时分析和智能决策需要的支撑体系[1].因此,数据断层的研究具有与信息系统的建设、智能决策支撑体系的架构同样重要的地位和应用价值.然而,从现有的文献资料来看,各行各业都仅仅是在文字描述中提及企业信息化、电子政务、金融投资、医学成像等领域存在着各种形式的"数据断层(data faultage)",却很少有从定义、体系、算法、模型的角度给予详细描述的.文献[2]提出基于数据仓库的环境,通过借鉴地质学领域断层的理论体系,将其运用到数据挖掘领域,并通过对其重定义、再分析,使断层在数据挖掘领域得到了很好的呈现和引申,并对宏观环境和微观环境下的数据断层进行了定量描述,给出了相关算法和模型.

在微观层面,数据对象是以一定的主题和存储模式进行聚集的[3],通过满足用户特定的需求为目标来分析数据对象,是数据挖掘和智能决策支持的关键[4].当数据总量并不大时,可以人为地对影响用户决策的数据对象进行简单处理;但是,当数据总量超出人们处理的范围,这种人为处理就很难达到预期效果[5-6].数据对象聚集以后,不符合要求的数据对象就会以一定规律汇聚形成微观数据断层(micro data faultage),这些断层一方面占用大量的内存空间,另一方面给后续的数据和结果分析带来一定的干扰.所以,为了提高分析结果的准确性,务必需要对断层数据作相应的检测、分析和处理,从而为精确化决策提供最大化支持.

根据数据断层所依赖的数据类型可以将数据断层分为数据显断层(dominant data faultage)和数据隐断层(tacit data faultage)两大类[2].本文借鉴地质断层领域中的概念和方法,针对数据显断层和数据隐断层引入相应的处理方法,采用缝隙(crack)检测与数据聚合(data aggregation)的方法来处理数据显断层,采用隐断层检测和数据熔合(data fusion solution)来处理数据隐断层,对上海"动感101"音乐电台的日志数据进行了数据显断层和数据隐断层的分析,达到了预期目的.

1 微观数据断层的基本定义

定义1 微观数据断层 专门用于描述同类型数据对象集合中或者同一数据对象集合中各个数据对象之间随着各种结构(各个数据对象的存放形式和构造)、成分(各个数据对象的内涵)、数据关系(各个数据对象之间的关系密切程度)等因素变化而变化的相关性质表象,称为微观数据断层.

定义2 数据显断层 主要用于描述大部分存在于数据对象集合与数据对象集合之间,受到数据对象集合的结构、主题、时效等因素影响而发生变化的较为明显的微观数据断层称为数据显断层.

定义3 数据隐断层 主要用于描述大部分存在于数据对象集合内部,受到数据对象集合内部的结构、成分、数据关系等因素影响而发生变化的较为隐秘的微观数据断层称为数据隐断层.

通过分析数据断层的表现形式所涵盖的内在特性,数据断层都存在必然性、易变性、非均质性(inhomogeneity)等特性[7].数据对象集合与集合之间以及对象集合本身会存在数据断层现象,这不仅具备了数据断层的必然性、易变性和非均质性等基

本性质，还具有客观存在性（existence）、隐秘性（mystery）和规律性（regularity）等一些特殊性质.

客观存在性表现在任何数据对象内部或多或少存在与主题无关的数据，即噪声数据[8]；隐蔽性指那些看上去与文章大意相同的数据对象[9]，它们的不一致性表现不明显，很难被发现；规律性主要是指数据显断层和数据隐断层在一定操作规程的约束下，某种程度上能够相互转化.

2　数据显断层

正确的决策支持需要数据对象的准确性来支撑，必然对数据对象集合的产生和存储提出了高质量的要求，这就要求数据对象本身应该具有准确性、可靠性、完整性和一致性的基本特征[10]. 但现实的数据却不尽如此，由于数据量本身的庞大和杂糅性，必然导致一系列数据问题，包括：不一致值、同一实体的多种描述、拼写问题、简写、打印错误、缺失值、未遵守引用完整性规则、不合法值等[11-12].

2.1　缝隙

本文参考基于图像分析的砂岩孔隙网络定量分析的模型[13]，将结合地质学原理，采用缝隙的概念来描述数据显断层中的问题数据对象.

定义 4　缝隙 数据对象中的缝隙数据包括与主题无关的所有数据对象，重点有空白数据对象、噪声数据对象、重复数据对象等，均定义为数据对象集合中的缝隙.

定义 5　缝隙度（crack degrees）针对某一特定主题，数据对象集合中存在的缝隙数量与数据对象集合所含有的所有数据对象数量之间的比值，称为该特定主题下的缝隙度.

在某一特定主题环境中，数据对象集合中数据对象之间的紧密程度可以通过缝隙度进行反映，即缝隙度越大，数据对象的紧密程度越低. 因此，针对不同的主题，同一数据对象集合的缝隙度并不一致.

定义 6　总缝隙度（total crack degrees）数据对象聚合中的所有与主题无关的数据与总体数据的比值，称为特定主题下数据对象集合的总缝隙度.

通常，总缝隙度可以表示为

$$P_{\mathrm{t}} = \frac{\sum d_e}{d_{\mathrm{r}}} \times 100\% \tag{1}$$

式中：$\sum d_e$ 指与特定主题无关的数据量；d_{r} 指数据总量；P_{t} 是总缝隙度.

为了满足某个特定的用户需求，有时候只需要对数据对象集合的局部进行分析处理即可，而并不需要对整个数据对象集合进行操作.

定义 7　有效缝隙度（valid crack degrees）　有效缝隙度指的是在具体的主题下用户选择的数据对象的缝隙与总的数据对象的比值.

通常，有效缝隙度可以表示为

$$P_{\mathrm{e}} = \frac{\sum d_i}{d_{\mathrm{r}}} \times 100\% \tag{2}$$

式中：$\sum d_i$ 是用户所选择的数据对象中存在的缝隙数量；P_{e} 是有效缝隙度.

为了表述简单，如非特别指出，本文均指有效缝隙度. 由于缝隙产生分先后次序，故将缝隙又划分为既生缝隙和再生缝隙两种类型.

定义 8　既生缝隙（existing crack）　既生缝隙指在数据集合形成的初始就存在的各类缝隙.

任何数据对象集合构建的时候，都必然存在构建此数据对象集合的目标主题，因此，此时数据对象集合内部所存在的缝隙都是既生缝隙. 随着时间的变化，数据对象集合本身的目标主题也会变化，那么将会导致缝隙的消失或者增大，还可能产生新缝隙.

定义 9　再生缝隙（newly crack）　数据对象集合随着时间、主题目标等变化产生的不同于既生缝隙的缝隙称为再生缝隙.

数据对象中可能存在既生缝隙和再生缝隙的混杂，因为数据的更新会造成原本只存在既生缝隙的数据对象又增加新的再生缝隙的情况. 根据数据挖掘的需要，应尽可能降低缝隙度，从而减小缝隙对分析数据的影响.

2.2　缝隙类型

在数据对象集合中，缝隙的类型主要有 3 种：

a. 缺失数据.

各种应用信息系统经过长期运行所积累下来的数据，均存在不同程度的缺失现象，即包含很多缺失数据. 从数据库的观点而言，如果某条记录中的某个属性存在空缺值，这就是缺失数据[14]. 按照数据对象集合中数据对象的不完整性，缺失值从缺失的分布来讲可以分为完全随机缺失、随机缺失和完全非随机缺失等[15].

b. 噪声数据.

简单来说，数据噪声指在一组数据中无法解释的数据变动，就是一些不和其他数据相一致的数据[16-17]. 噪声数据有可能是数据对象集合中存在的

错误数据,也有可能是数据对象集合形成过程中记录下来的随机错误或者偏差[14],也有可能是完全不相关的数据对象或者是没有实际意义的数据对象[18].

噪声数据的产生通常有3类原因[19],即:不同数据对象集合由数据集成操作引发;数据对象本身随着某一因素的改变而自然发生;在数据测量和采集过程中由人为填写错误或者仪器故障等原因引起.

c. 重复数据.

重复数据是指对同一实体对象的多重表述,或者说在数据结构或数据表示上对同一实体对象的不同表达[20].任意的数据对象集合中都有重复数据记录或数据属性的可能性,因此,几乎所有的信息系统都可能存在着信息重复和冗余的现象[21].通常,检测重复数据的目的是从数据对象集合中识别出那些在表现形式上不同但又表示同一实体对象的记录[14,22].

2.3 缝隙检测

缝隙检测的目的有两个,首先是判断有没有缝隙,再者,若有缝隙,鉴别缝隙大小和位置.现实状态下,无论是缺失数据、噪声数据、重复数据等具有代表性的缝隙,还是其他特殊类型的缝隙,都会影响正常的数据预处理工作和数据分析工作,因此,这类问题十分值得研究和分析[15,19,22].

进行正常的数据预处理工作就是对数据对象集合进行筛选操作,尽可能多地剔除缝隙[6].缝隙检测的常用方法包括:人工神经网络法[23]、统计判别法[24]、支持向量机法[25]等.

本文借鉴基于能量的概念[26]和基于层次聚类的方法[27],提出了基于能量和层次聚类的缝隙检测(crack detection on energy and hierarchy clustering)方法.为了定量化进行缝隙查找,引入相关能的概念进行描述,缝隙的检测问题进一步转化成在数据对象集合中通过相关能计算获取较大能值的数据对象过程.

假定 d 维数据空间 F^d 中有 x 和 y 两个数据对象.当两个数据对象一致($x=y$)时,合并两个数据对象;若两个数据对象不一致($x\neq y$)时,则计算其相关能值的公式为

$$E_c(x,y)=\frac{tE_p(x)E_p(y)}{k^2\sqrt{(x-y)^2}} \tag{3}$$

式中:k 是常系数;t 是不确定系数;$E_p(x)$ 表示数据对象 x 的势能[7].

算法 1 基于能量和层次聚类的缝隙检测

Step 1 根据用户需求,选定特定主题,并将数据对象集合分为 num 个数据分区.

Step 2 对数据分区进行层次聚类操作.

a. 在数据分区中,先假定各数据对象相互独立,然后使用式(3)计算类与类之间的相关能;

b. 设定类合并的条件(根据具体情况设定);

c. 将相关能由大到小排序,将相关能最大(满足类合并条件)的两个类合并成一个类;

d. 在数据分区中对新生成的类与其他类重新计算相关能;

e. 重复步骤 c 和 d,直至无法再进行类归并为止.

Step 3 对数据分区进行层次聚类处理后,若数据分区不存在异常,则将其从原数据分区中减除.

Step 4 重复执行 Step 2 和 Step 3,完成 num 个数据分区的所有检测.

如果数据分区中存在异常数据对象,则单列该数据分区,以便单独分析和处理.

基于能量和层次聚类的缝隙检测方法具有明显的3大优势:(a)缝隙检测的基本单位是数据分区,不需要将所有的数据全部调入内存,只需将要检测的数据分区调入内存即可;(b)计算数据对象间相关能的计算量取决于数据分区中的数据对象总数,由于分区内的数据对象总数相对较少,因此计算量大大减少;(c)数据对象集合中不存在缝隙的数据分区不需要进行计算,因此整体的计算量也相应减少.

缝隙检测的目的并不在于通过检测来消除它,而是在于发现缝隙与数据断层之间主从关系所折射出来的数据对象之间变化的规律.如果这些数据对象是符合用户最终需求的,那么缝隙检测的分析结果就是数据断层的分析结果;但是如果这些数据对象并不符合用户的最终需求,则需要进一步的数据聚合操作才能避免数据断层的危害.

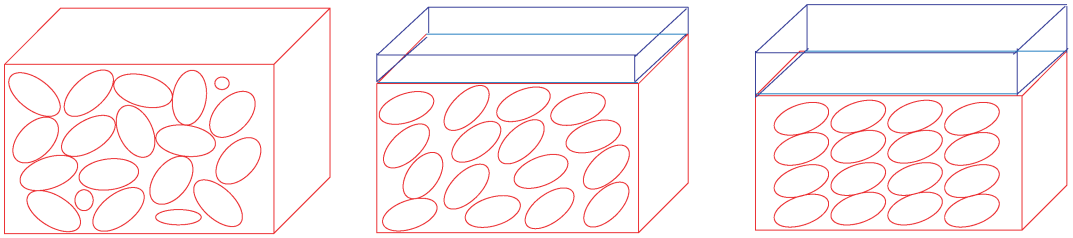
2.4 数据聚合

定义 10 数据聚合 为了防止缝隙的存在对后续数据分析造成负面影响而进行的各种操作,就是数据聚合.

数据聚合通过处理数据对象集合中存在的缝隙,降低缝隙对数据分析的不准确性造成的影响.数据聚合操作有:对空白数据对象的处理、对格式不一致数据对象的处理、对重复数据对象的处理等.数据聚合是微观数据断层研究中的关键步骤,聚合数据对象集合中存在的缝隙是为后续的数据分析与优化决策服务的.

数据聚合的示意图如图1所示.图1(a)为数据

对象集合的原始状态图，集合中存在明显的缝隙且集合中数据对象排列混乱；通过数据聚合初步处理后的情形如图 1(b)所示，缝隙已经被消除，数据排列趋于半结构化，数据集合体积总体减小；当对数据对象进一步聚合操作后，就得到了更加统一化、规整化和结构化的数据对象排列，如图 1(c)所示.



(a) 数据聚合原始状态　　(b) 初步数据聚合　　(c) 完全数据聚合

图 1　数据聚合示意图

Fig.1　Diagram of data aggregation

数据聚合的程度可以使用聚合率来表示

$$C = \frac{d_c}{d_t} \times 100\% \tag{4}$$

式中：d_c 表示通过数据聚合操作删除的数据对象数量；d_t 表示数据对象集合中原有的数据对象数量. 在实际应用中，聚合率反映数据对象集合对用户具体需求的满足程度，即聚合率越低，符合要求的数据对象越多.

在具体应用中，数据质量、缝隙形态、聚合方法都会对数据聚合后的效果产生重要的影响. 通常，数据对象结构越良好，数值规范越准确，数据聚合的效果越好；缝隙类型越单一，缝隙内容越关联，数据聚合的操作越简单.

基于能量和层次聚类的缝隙检测方法能够明确缝隙存在的位置以及缝隙具体的类型，采用不同的数据聚合方法可以处理不同类型的缝隙.

a. 聚合缺失数据.

聚合缺失数据虽然操作简单，但是选择过程比较繁琐，在具体应用中主要有删除法和填充法两种可供选择. 对于缺失值的处理，从总体上来说分为删除存在缺失值的个案[28]和缺失值插补[29-30]. 通常采用的填充法包括均值、同类均值、缺失值的预测、全局常量和人工等 5 种.

均值法是使用属性的均值来填充缺失的数据项，这对于数据缺失现象比较均匀的情况特别适用；同类均值法是针对某种特定类型的数据发生缺失，将数据分类后以特定类型数据的均值作为缺失数据进行填充；缺失值的预测法就是采用一些数据分析方法对缺失值进行合理预测，选择预测值最大的值进行填充；全局常量法就是对每一个缺失值采用固定的全局常量来填充；在数据缺失情况比较严重的时候，通常采用人工操作，但其对于填充人的主观依赖性过于明显，存在较大的准确性偏差.

b. 聚合噪声数据.

聚合噪声数据通常采用分箱法[31]、聚类法[32]、回归法[33]等诸多常规方法.

分箱法是对数据周围的数据进行分析，按照不同的取值方法来平滑数据的值；聚类法是将数据对象集合中的每个数据视为对象，按照相似程度把对象划分到一个个类中，提高类内对象的相似度，降低类间的相似度；回归法是尽可能地利用称为线性回归方程的最小平方函数对一个或多个自变量和因变量之间关系进行建模，从而推断出不合理数据可能对应的正确值.

c. 聚合重复数据.

重复数据的处理通常采用把重复数据直接从数据对象集合中删除的方法来实现消除数据冗余的目的. 这不仅有利于降低数据对象的空间占有度，也有利于优化数据对象的存储容量. 删除重复数据主要有固定块删除法与可变块删除法两种[34].

固定块删除法就是将数据划分成长度固定的部分或块，然后执行删除操作；可变块删除法就是根据一个大小可变的滑动窗口对存在的重复数据进行删除操作.

3　数据隐断层

不同的数据对象其存储方式也不同. 正是由于存在数据对象维度的差异性，需要引入数据对象集合的数据密度(data density of set)这一概念来描述数据对象的分布情况[7].

定义 11　数据对象集合的数据有效密度 假设

将整个数据空间分为 k 个数据对象，假定某个数据对象由$\{a_1,a_2,\cdots,a_n\}$组成，$(n>0)$，$a_{\max}$和 $a_{\min}$是数据对象中的最大和最小值，$a_{\max}\neq a_{\min}$，则数据对象集合 K 的数据密度定义为数据对象集合K 中所有数据值之和与其最大值和最小值的差值之比

$$d_K=\frac{\sum_{j=1}^{n}a_j}{a_{\max}-a_{\min}} \tag{5}$$

在定义 11 的基础上，进一步定义数据对象集合的数据均值和数据对象集合的断层概率.

定义 12 数据对象集合的数据均值 数据对象集合 K 中所有数据值之和与其数据对象的数量之比定义为数据对象集合K 的数据均值.

$$E(K)=\frac{\sum_{j=1}^{n}a_j}{n} \tag{6}$$

定义 13 数据对象集合的断层概率 对于数据对象集合 K 中任一数据值 a_j 所对应的数据对象 K_{a_j}，定义其在数据对象集合 K 中发生数据断层的概率 $p_{K_{a_j}}$、数据密度 d_K 和数据均值$E(K)$的关系

$$p_{K_{a_j}}=\frac{|a_j-E(K)|}{nd_KE(K)}\times 100\% \tag{7}$$

3.1 隐断层侦测

隐断层可能会因为用户需求不同而不同，所以数据对象集合中某些隐断层可以很容易地被侦测出来，例如采用观察法或总结归纳法等，但是有些隐断层的发现并不简单. 通常，需明确数据隐断层定义，并计算出数据对象集合的信息熵，并将其与相应阈值进行比较才能得出结果. 但是，从定义法来侦测隐断层也存在着一些缺陷，例如隐断层的状态不能依据定义判断出来. 因此，本文采用基于定义与数据密度的断层侦测(faultage detection on definition and data density)方法来对隐断层的位置与程度进行确定.

算法 2 基于定义与数据有效密度的断层检测

Step 1 根据用户的主题需求，将数据空间划分为 k 个数据对象集合.

Step 2 计算 k 个数据对象集合各自的信息熵 $H(X_i)=-\sum_i p_i\mathrm{lb}\,p_i$.

Step 3 再次判断用户的实际要求，明确阈值 T.

Step 4 分别对各个数据对象集合的信息熵 $H(X_i)$与阈值 T 进行比较. 如果信息熵 $H(X_i)$小于阈值 T，则判定数据对象集合存在隐断层，否则不存在.

Step 5 对于已侦测出存在隐断层的数据对象集合，则需要根据定义 11～13 计算出数据对象集合的断层概率，判断隐断层存在的确切位置.

基于定义与数据有效密度的断层检测方法中确定阈值 T 是至关重要的一步，针对不同的应用情况，其确定的方法也不尽相同，本文不再赘述. 不同的隐断层数据需要不同的处理方法来配套. 若隐断层数据可以为用户决策所服务，对于这类隐断层数据则需要将其提取出来进行进一步分析；若隐断层数据是无意义的，那么采用数据熔合进行处理即可.

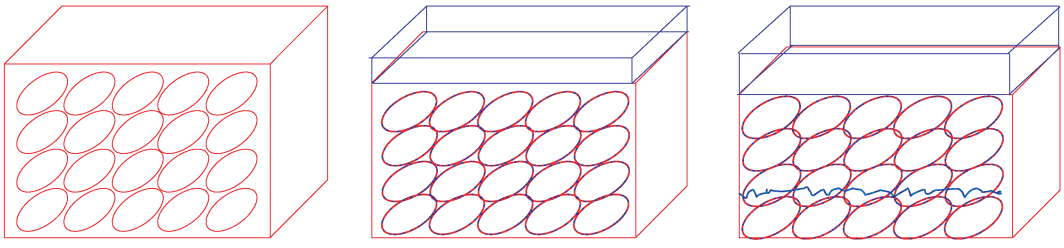
3.2 数据熔合

隐断层的存在并无好坏之分. 隐断层数据的处理方式依据用户具体需要确定，可以进行分析，也可以进行消除. 本文引入数据熔合来消除隐断层数据.

定义 14 数据熔合 对不满足用户需要的数据进行处理进而获得有价值的信息的过程定义为数据熔合.

数据熔合的过程与数据聚合的过程相类似，也主要由 3 大步骤组成(如图 2 所示). 但是在数据熔合中，非缝隙数据对象在熔合处理后自身会发生变化，包括隐断层数据对象的隐藏或转化.

经过数据熔合初步处理后，边缘数据对象排列得更加紧凑，同时减少了数据对象所占用的空间. 对数据对象集合进行大规模数据熔合后，隐断层数据



(a) 数据熔合原始状态　(b) 初步数据熔合　(c) 完全数据熔合

图 2 数据熔合示意图

Fig. 2 Diagram of data fusion solution

得到消除,数据对象按照相同主题进行聚集,数据空间得到充分利用.数据熔合具体操作方法需要依据具体的应用情况而采用合适的方法,大致分为 3 种[7].

a. 人工处理.

传统的人工处理方法主要是借助类似 SPSS,SAS 等传统的数据分析软件工具来进行.这种处理方法主要适用于数据量不大的情况,当数据量较大时,这种方法效率十分低下,难以有效处理任务.

b. 程序处理.

设计一个程序对存在断层的数据对象集合进行分析处理,与人工处理方法相比,程序处理方法效率相对提高,可以处理较大容量的数据对象集合.但是,这种处理方法灵活性较差,程序严格依赖于数据结构与属性,不具有通用性,因此这种方式只在数据对象形式固定的情况下比较适用.

c. 领域无关处理.

采用与领域无关的数据熔合方式,表现形式较多,通常有属性约简[35]和数据规范化[36]两种.

通常,数据对象集合中可能包括成百上千的数据属性,但是数据挖掘只需涉及数据对象集合的极少一部分相关属性并进行相应处理,从而降低整体的数据数量.

另外,在同一个数据属性中,其数值波动的范围可能十分巨大,这种情况会显著降低一些数据挖掘算法的执行性能和效率.因此,需要利用数据的规范化方法,尽可能把数据的属性值映射到一个幅度不大的区间中[29].

4　针对电台收听数据的断层分析实验

本文以上海"动感 101"频道的移动客户端用户访问日志数据(以下简称"电台数据")为研究对象,对日志数据中可能存在的数据断层进行分析和处理.

本文的实验数据是"动感 101"电台移动客户端 2013 年 4 月 8 日 0 时到 2013 年 4 月 14 日 24 时的访问日志数据,来自源 IP 地址为 222.＊＊＊.＊＊＊.167,222.＊＊＊.＊＊＊.207,222.＊＊＊.＊＊＊.208 的服务器,总文件大小约为 3.1 G.一条记录为一个切片,且一个切片代表用户进行连续 10 s 的访问操作.鉴于原始电台数据的属性较多,根据研究者实际的需求,只选择了如表 1 所示的 5 个数据属性进行实验.

表 1　选择的 5 个数据属性

Tab.1　Selected data attributes

数据属性	含义
date	用户访问日期
time	用户访问时间
ts	下载的切片流
ip	用户 IP 地址
mobile	用户设备参数

4.1　数据显断层分析和处理

根据算法 1 的检测结果,"动感 101"电台数据的缝隙类型主要有缺失数据、噪声数据和重复数据.

缺失数据的主要表现为数据属性 ts 是空的,这表示用户并没有下载切片流;噪声数据的主要表现为数据属性 mobile 显示 LiveRadioEncoder 或者 ChinaCache 等字样,表示这些记录是由内部编码器向服务器发送音频切片文件所产生的访问记录,而非真正的日志访问记录;重复数据的主要表现是存在一些每个数据属性都相同的记录,这类记录大多是访问页面的信息,并没有下载任何的切片流信息.

3 台服务器一周内电台数据所表现的 3 种缝隙数据统计信息如表 2 所示.

表 2　一周缝隙数据统计

Tab.2　Statistics of the crack data within a week

日期	缺失数据	噪声数据	重复数据
2013-04-08	28 291	123 285	70 306
2013-04-09	21 456	108 336	71 043
2013-04-10	28 278	123 588	67 110
2013-04-11	28 256	123 536	81 362
2013-04-12	28 289	123 785	68 720
2013-04-13	28 266	123 584	72 035
2013-04-14	28 264	123 040	69 377

图 3(见下页)显示的是对电台数据缝隙度的统计信息,电台数据缝隙度基本在6.81%～10.98%之间波动.如图 3 所示,4 月 8 日(周一)到 4 月 11 日(周四)的缝隙度相对比较稳定,而 4 月 12 日(周五)到 4 月 14 日(周日)的缝隙度明显升高,可以猜测是由于用户从周五开始到周日对电台的收听明显减少,但应用程序总量却没变,从而导致缝隙度变大.

对于实验中提到的 3 种主要类型缝隙数据,本实验都分别进行了数据聚合操作,根据特定的情况,主要采取删除法.

4.2　数据隐断层分析和处理

完成数据显断层的分析和处理后,按照算法 2

对“动感101”电台数据进行检测.

为了拟合用户的实际需求,将“动感101”电台数据中国外IP地址访问的数据记录进行了删除,同时按照天为单位,对每天重复的国内IP地址访问的数据记录也进行了留一条的删除操作,然后才对相应遗留的用户数进行统计,并按照地理位置的划分将数据对象划分到9个地区的集合中.

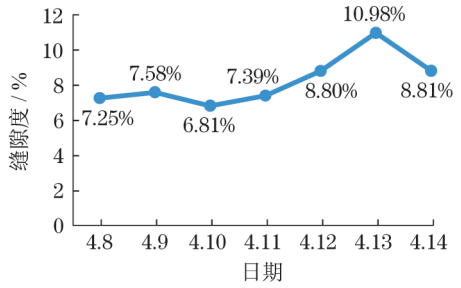


图3 电台数据的缝隙度

Fig.3 Crack degrees of “Dynamic 101”

图4显示了按照地理位置划分的收听人数分布情况.由于“动感101”是属于上海地区的电台,上海的用户固然很多,再加上临近上海的江苏和浙江的用户人数,使得东部沿海地区的用户数远远高于其他地区的用户数,从而导致东部沿海地区与其他地区在用户数量上产生了断层.

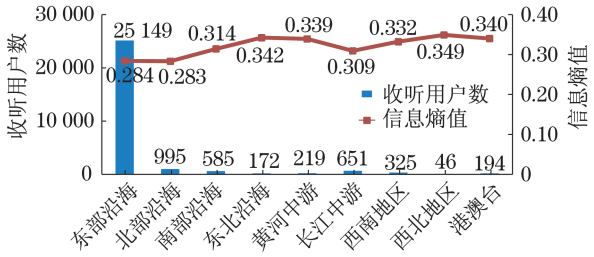


图4 收听用户区域分布及其信息熵

Fig.4 Listener distributions and its entropy of information

同时,图4显示了9个地区集合的信息熵情况,大部分分布在0.30～0.35之间.9个地区的信息熵,其中有7个介于0.30～0.35之间,只有东部沿海和北部沿海低于0.30.如果根据实际需求将阈值定为0.30,那么存在数据断层的就是东部沿海和北部沿海地区.通常信息熵越大,数据断层概率相对小,数据断层数量和规模也相对小.

忽略其他不存在数据断层的地区,重点看一下存在数据断层的东部沿海和北部沿海,再根据定义11～13计算各个地区中省市的断层概率,如表3所示.在东部沿海地区,考虑到在用户数方面上海是浙江和江苏的接近9倍,导致上海的数据隐断层概率高达0.337 5.在北部沿海地区,北京的用户数是天津、河北和山东的4倍左右,导致北京的数据隐断层概率高达0.160 4.

断层概率的确定,对于数据隐断层的确切位置给予了明确,并进一步揭示了隐藏在数据背后的影响力问题.固然,IP地址标识为上海的电台数据本来就因为听众数量的关系占据着主导地位,但是通过断层概率分析,IP地址标识为北京的电台数据尽管在人数上与浙江、江苏相差较远,但是其在区域性主导地位方面仍然具有重大影响力.

表3 东部沿海和北部沿海地区的断层概率

Tab.3 Faultage probability of east coast and north coast

区域	省市	听众数量	断层概率
东部沿海	上海	19 815	0.337 5
	浙江	2 330	0.169 8
	江苏	2 402	0.167 7
北部沿海	北京	571	0.160 4
	天津	157	0.041 6
	河北	128	0.055 7
	山东	113	0.063 0

5 结 论

作为实现管理决策变革的三要素之一,建立完善能够和谐平衡各个信息系统之间数据断层的机制一直以来是智能决策支持系统建设领域重要的研究内容,并在众多领域得到初步的实践与应用[2].本文选择数据挖掘领域的“数据断层”概念,分别探讨和分析了数据显断层领域的缝隙现象和减少缝隙的数据聚合方法,以及数据隐断层领域的隐断层现象和减少隐断层的数据熔合方法,并就具体案例的日志数据分析,对蕴含其中的数据显断层和隐断层进行了细节化论述.

整个数据显断层和隐断层的分析处理过程对于引导用户根据自身要求得到有用信息并进行智能决策都具有良好的指导作用[37-39].因此,进一步深化数据断层的研究,有利于强化数据挖掘技术对于智能决策支持领域的支撑作用,有利于创新数据挖掘领域的研究.

参考文献:

[1] 赵栩.智慧的未来供应链趋势展望[EB/OL].(2011-

10－15)[2015－07－22]. http://articles.e-works.net.cn/scm/article91405.htm.

[2] 夏骄雄，汪晶玲，严琛琼，等. 数据断层现象的研究[J]. 计算机应用与软件，2013，30(8)：9－13.

[3] FLEISCHMANN A，SCHMIDT W，STARY C. Subject-oriented development of federated systems——a methodological approach[C]// Proceedings of the 2014 40th EUROMICRO Conference on Software Engineering and Advanced Applications. Verona，Italy：IEEE Computer Society，2014：199－206.

[4] SHARMA D，SHADABI F. Multi-agents based data mining for intelligent decision support systems[C]// Proceedings of the 2014 2nd International Conference on Systems and Informatics. Shanghai：IEEE Computer Society，2014：241－245.

[5] ZHENG L，DHU W，MIN Y. Raw wind data preprocessing：a data-mining approach [J]. IEEE Transactions on Sustainable Energy，2015，6(1)：11－19.

[6] 夏骄雄. 数据资源的聚类预处理[M]. 上海：上海科学普及出版社，2011：11.

[7] 汪晶玲. 数据资源中数据断层现象的研究[D]. 上海：上海大学，2013.

[8] MILLER L D，SOH L K. Cluster-based boosting[J]. IEEE Transactions on Knowledge and Data Engineering，2015，27(6)：1491－1504.

[9] ZHANG H，CHEN G，OOI B C，et al. In-memory big data management and processing：a survey[J]. IEEE Transactions on Knowledge and Data Engineering，2015，27(7)：1920－1948.

[10] ALGHASSI A，PERINPANAYAGAM S，SAMIE M，et al. Computationally efficient，real-time，and embeddable prognostic techniques for power electronics[J]. IEEE Transactions on Power Electronics，2015，30(5)：2623－2634.

[11] QIU F D，WU F，CHEN G H. Privacy and quality preserving multimedia data aggregation for participatory sensing systems[J]. IEEE Transactions on Mobile Computing，2015，14(6)：1287－1300.

[12] SHORE J E. Relative entropy，probabilistic inference，and AI [C] // Proceedings of the First Annual Conference on Uncertainty in Artificial Intelligence. Los Angeles：Elsevier，1988：211－216.

[13] BERREZUETA E，GONZÁLEZ-MENÉNDEZ L，ORDÓÑEZ-CASADO B，et al. Pore network quantification of sandstones under experimental CO_2 injection using image analysis [J]. Computers & Geosciences，2015，77：97－110.

[14] SSALI G，MARWALA T. Computational intelligence and decision trees for missing data estimation[C]// Proceedings of the International Joint Conference on Neural Networks. Hong Kong：IEEE，2008：201－207.

[15] VATEEKUL P，SARINNAPAKORN K. Tree-based approach to missing data imputation[C]// Proceedings of the IEEE International Conference on Data Mining Workshops. Miami，Florida：IEEE，2009：70－75.

[16] JIN W，TUNG A K H，HAN J W. Mining top-n local outliers in large databases[C]// Proceedings of 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco：ACM Press，2001：293－298.

[17] WANG Y T，FENG H Y. Outlier detection for scanned point clouds using majority voting[J]. Computer-Aided Design，2015，62：31－43.

[18] XIONG H，PANDEY G，STEINBACH M，et al. Enhancing data analysis with noise removal[J]. IEEE Transactions on Knowledge and Data Engineering，2006，18(3)：304－319.

[19] CHOI J，KIM K E. Hierarchical bayesian inverse reinforcement learning [J]. IEEE Transactions on Cybernetics，2015，45(4)：793－805.

[20] FERRO A，GIUGNO R，PUGLISI P L，et al. An efficient duplicate record detection using q-grams array inverted index [C] // Proceedings of the 12th International Conference on Data Warehousing and Knowledge Discovery. Berlin Heidelberg：Springer，2010：309－323.

[21] BILENKO M，MOONEY R J. Adaptive duplicate detection using learnable string similarity measures [C] // Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC：ACM Press，2003：39－48.

[22] ELMAGARMID A K，IPEIROTIS P G，VERYKIOS V S. Duplicate record detection：a survey [J]. IEEE Transactions on Knowledge and Data Engineering，2007，19(1)：1－16.

[23] HONG Y S T，ROSEN M R，BHAMIDIMARRI R. Analysis of a municipal wastewater treatment plant using a neural network-based pattern analysis[J]. Water Research，2003，37(7)：1608－1618.

[24] NGUYEN T，PHUNG D，DAO B，et al. Affective and content analysis of online depression communities[J]. IEEE Transactions on Affective Computing，2014，5(3)：217－226.

[25] ZENG G M，LI X D，JIANG R，et al. Fault diagnosis of WWTP based on improved support vector machine [J]. Environmental Engineering Science，2006，23(6)：1044－1054.

[26] JOHNSON S. Optimization strategy of energy-description and collision-description clustering[J]. Journal of Information and Computational Science, 2011,8(8):1251-1260.

[27] BIAN M J,XIA J X,XU J. Database preprocessing with AHP[C]// Proceedings of the 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery. Yantai:IEEE Press,2010:2805-2809.

[28] MAYFIELD C S, NEVILLE J, PRABHAKAR S. ERACER:a database approach for statistical inference and data cleaning[C]// Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. Indianapolis,Indiana:ACM Press,2010:75-86.

[29] KOTSIANTIS S B, KANELLOPOULOS D, PINTELAS P E. Data preprocessing for supervised learning[J]. International Journal of Computer Science, 2006, 1(2):111-117.

[30] BRUHA I,FRANEK F. Comparison of various routines for unknown attribute value processing: the covering paradigm[J]. International Journal of Pattern Recognition and Artificial Intelligence, 1996, 10(8): 939-955.

[31] LUANGPAIBOON P, CHINDA K. Computer-based management of interactive data transformation systems using Taguchi's robust parameter design[J]. International Journal of Computer Integrated Manufacturing,2015,28(10):1030-1045.

[32] BUADHACHÁIN S Ó, PROVAN G. An efficient decentralized clustering algorithm for aggregation of noisy multi-mean data[J]. Journal of Heuristics,2015, 21(2):301-328.

[33] GHASSABEH Y A, RUDZICZ F. Noisy source vector quantization using kernel regression[J]. IEEE Transactions on Communications, 2014, 62(11): 3825-3834.

[34] WAGNER C,MILLER S,GARIBALDI J M,et al. From interval-valued data to general type-2 fuzzy sets[J]. IEEE Transactions on Fuzzy Systems, 2015, 23(2): 248-269.

[35] 葛浩,李龙澍,杨传健.基于相对分辨能力的属性约简算法[J].系统工程理论与实践,2015,35(6):1595-1603.

[36] XU Y W,HOU C H,YAN S F,et al. Fuzzy statistical normalization CFAR detector for non-rayleigh data[J]. IEEE Transactions on Aerospace and Electronic Systems,2015,51(1):383-396.

[37] HSU D, JOHNSON S, HUI M. Data faultage in data resource[J]. International Journal of Database Theory and Application,2015,8(6):271-284.

[38] 徐俊,夏骄雄,周时强.数据断层分析在广播电台数据处理中的应用[J].计算机应用与软件,2016,33(9):38-42,158.

[39] 夏骄雄,刘政,刘绪彬,等.基于快速应用开发的功能点增量迭代模型[J].上海理工大学学报,2014,36(6):578-584.

(编辑:丁红艺)