doi:10.11959/j.issn.1000-436x.2016068

基于 HBase 数据分类的压缩策略选择方法

王海艳[1,2]，伏彩航[1,2]
（1. 南京邮电大学计算机学院，江苏 南京 210023；2. 南京邮电大学江苏省无线传感网高技术研究重点实验室，江苏 南京 210003）

摘 要：为解决现有的 HBase 数据压缩策略选择方法未考虑数据的冷热性，以及在选择过程中存在片面性和不可靠性的缺陷，提出了基于 HBase 数据分类的压缩策略选择方法。依据数据文件的访问频度将 HBase 数据划分为冷热数据，并限定具体的访问级别；在此基础上增加评估层，综合考虑基于相邻区和统计列的选择方法，提出基于数据访问级别的压缩策略选择方法。仿真实验及结果表明，提出的压缩策略选择方法不仅节省了存储空间，还大大提高了数据查询的性能。
关键词：数据压缩；HBase；压缩策略选择方法；冷热数据
中图分类号：TP301.6 **文献标识码：**A

Compression strategies selection method based on classification of HBase data

WANG Hai-yan[1,2], FU Cai-hang[1,2]
(1. School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
2. Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Most of the current compression strategies selection methods for HBase data did not consider whether the data was cold or hot. Besides, problem of incompleteness and unreliability existed during selection process. To address the problems above, a compression strategies selection method based on classification of HBase data was put forward. HBase data was classified into cold and hot data according to the access frequency of each data file and an access level would be designated to each file. On this base, an evaluation layer was added and a compression strategies selection method based on access level with integration of neighbor sector and statistic column based selection methods. Simulation experiments and results demonstrate that the proposed compression strategies selection method based on classification of HBase data can not only save storage space but also greatly improve the query performance of HBase data.
Key words: data compression, HBase, compression strategies selection method, cold and hot data

1 引言

据互联网数据中心（IDC）多年的研究结果表明，全球数据量大约每 2 年翻一番，每年产生的数据量按指数增长，数据增速符合摩尔定律，预计到 2020 年，全球总数据量将达到 35 ZB[1]。如何对这些海量数据进行有效分析和处理已成为大数据实时分析应用的一个重要环节。HBase 正是实现数据实时分析的一个主要技术。目前，关于 HBase 数据库压缩方法的研究工作取得了一定的成果，已有的 HBase 压缩技术根据数据特征信息选择对应的算法进行压缩，提出了基于相邻参照区和基于统计列等的压缩策略选择方法。然而，不断增长的海量数据给数据压缩存储带来了新的挑战，现有数据压缩主要存在以下问题。

现有压缩方法未考虑数据的冷热分类问题，冷热数据采用相同的压缩方法，导致存储成本较高。以往的基于 HBase 的压缩方法直接根据数据的不同分布特征来选择对应的压缩策略。然而，这种数据压缩的思路并未将数据的“冷热”性考虑在内。

收稿日期：2015-07-21；**修回日期：**2016-03-01
基金项目：国家自然科学基金资助项目（No.61201163）
Foundation Item: The National Natural Science Foundation of China(No.61201163)

HGST 的分析结果表明，只有 10%~15%的数据是被经常访问的，而其他全部是“冷数据”[2]。如果对那些不经常被访问的大量“冷数据”采用同样的压缩策略，必然导致存储空间极大浪费，存储效率低下。

以往压缩策略选择方法存在片面性和不可靠性的缺陷。目前，在压缩策略选择方法研究中，存在着两大方向：基于相邻参照区和基于统计列的压缩策略选择方法。前者是通过比较相邻区数据特征信息的相似性进行策略推荐，这种方法考虑的仅仅是数据的某个特征，即特征向量的某个分量，选择依据缺乏全面性和科学性。后者是从待压缩数据中抽取一定数量的样本并通过对样本进行学习得到先验知识，再对先验知识进行统计得到训练器——压缩算法选择的工具。但是，这种方法得到的训练器在压缩策略选择上是不够准确的。所以，需要对原有的压缩策略选择方法进行改进，以此来保证策略选择的准确性，提高数据压缩的效率。

针对以上问题，本文展开了深入研究，主要工作如下。

1) 提出一种基于访问频度的数据分类方法。根据一段时间内数据库文件的访问次数得到相应的访问频度，依据各数据文件的访问频度及相关阈值将数据文件划分为冷热数据并确定具体的访问级别。

2) 提出基于数据访问级别的压缩策略选择方法，定义了确定数据样本的抽样方法，针对原有的压缩策略选择方法中先验知识未必可靠的缺陷，通过添加评估层及时调整先验知识，并在基于相邻参照区和基于统计列选择方法的基础上设计出 HBase 数据压缩策略选择方法，优化存储成本。

2　相关工作

在数据压缩中，相对于行存储，列存储具有更大的压缩潜力，并且压缩的列数据比压缩的行数据更适合直接处理。基于列存储的压缩技术在大数据领域应用较多的是基于 HBase 的压缩技术，而 HBase 数据压缩存储方面的研究成果较少，本文最主要的工作就是基于数据分类的 HBase 压缩技术研究。

2.1　列存储压缩技术相关研究

列存储数据压缩技术在近年的研究中取得了一定的成果。在基于列存储的压缩策略选择方面，文献[3]提出了基于列的统计信息为参照信息,进而通过学习参照信息与当前区之间的相似性和差异性进行策略推荐；文献[4]提出了根据数据的分布情况，自适应地选择区的大小，灵活地进行区划分并且对每个区选择更合适的压缩策略。在列存储压缩策略改进方面，文献[5]提出了在列存储数据库中引入一些轻量级的压缩算法，以列属性值为一个编码单元进行数据压缩，减少了数据压缩和解压时间。文献[6]提出了一种高效的应用在无线传感网中的压缩算法 TinyPack，这种哈夫曼风格的压缩方案利用了时间局部性和三角洲压缩的特性，不仅有较高的压缩率，并提供了较好的带宽利用，而且也降低了延迟和能源开销。文献[7]提出了利用差向量的基于重排序和列式位填充的汉明距离（HDR-CBS-DV）方法，该方法可以和任何基于游程的编码技术同时使用从而提高压缩率，在使用基于游程的编码技术压缩之前，将该方法作用在对应的数据集上。

2.2　基于冷热数据分类的相关研究

目前，对冷热数据的分类研究主要在数据分级存储方面涉及较多。分级存储系统通常分为在线存储、近线存储和离线存储等三级存储方式。典型的数据分级算法是基于访问频率的分级算法，文献[8]提出了结合固定阈值法和高低水位法这 2 种典型的数据分级算法，设计了数据分级动态阈值法，通过前数次迁移阈值的计算对下一次的迁移阈值做一定的预测，从而达到每次迁移后都对阈值做出相应的修正的目的，提高了分级的精度。文献[9]对位于主存中的联机事务处理系统数据库（OLTP database）进行分析研究发现各类数据的访问频率是不同的，在此基础上提出了利用样本的方法收集数据的访问日志、进行线下分析从而预测数据的访问频度，并提出了 4 种预测效果较好的基于指数平滑法的预测算法。

2.3　基于 HBase 的压缩技术相关研究

HBase 支持的压缩算法有 Gzip 和 LZO 等。Gzip 使用 Deflate 算法进行数据压缩，Deflate 算法同时使用了 LZ77 算法和哈夫曼算法。LZO 是基于字典思想的压缩算法,LZO 的优点是解压速度快且不需要额外的内存。文献[10]提出了利用 HBase 存储语义传感网数据集的方法 LSD2H，并从减少存储空间和提高查询性能角度对 LZO、Gzip、Snappy 和不压缩的情况进行分析比较，得出了 LZO 是 LSD2H 方法中压缩效果最好的压缩算法。文献[11]提出了在 HBase 中对文件进行压缩处理时，先将其划分为多个大小相同的小文件，根据各个文件的访问频度对

小文件进行分类并进行压缩。文献[12]中提出的方法证明对于 LZO 压缩算法，相同访问频度文件的解压速度是 Gzip 的 2 倍，能够节省磁盘的读写工作，但压缩比不如 Gzip。文献[13]提出的基于朴素贝叶斯分类的压缩策略选择方法首先对样本进行学习，在此基础上对具体的数据选择对应的压缩策略，达到了较好的压缩效果。

3 相关概念介绍

定义 1 冷数据和热数据[2]。热数据是指被系统所实时使用且经常被访问的一类数据；冷数据是指那些很少或不再被系统所使用的一类数据。

定义 2 访问频度[8]。访问频度是指 HDFS 文件系统中一段时间内文件的平均访问次数，这里用 $frequency$ 表示，$frequency=\frac{C}{t}$。其中，C 指一段时间内数据文件的访问次数，t 是相应的时间段。

定义 3 HBase 数据文件访问级别。数据文件访问级别是指根据文件访问频度的不同而划分出的数据级别。数据访问级别是选择数据压缩算法要考虑的一个因素，热数据分为热点数据、次热点数据、活跃数据和不活跃数据 4 个级别；冷数据则对应沉默数据这一级别。

定义 4 数据块统计量[3]。数据块统计量是指根据数据块特征信息而量化出的一个向量，用 $\boldsymbol{q}$ 表示。$\boldsymbol{q}=\{q_1,q_2,q_3,q_4\}$，其中，$q_1$ 代表数据块中相同值所占的百分比，q_2 代表不同的记录数所占的百分比，q_3 代表相同值连续出现的平均长度，q_4 代表空值所占的百分比。

定义 5 数据块特征信息[3]。数据块特征信息是指数据块中数据分布的特征，是压缩算法选择的唯一依据，它包含数据块中所有属性值的个数、相同值的记录数、不同值的记录数、相同值连续出现的平均长度、空值的数目等 7 个分量，表示为：$\boldsymbol{T}=\{a_1,a_2,a_3,a_4,a_5,a_6,a_7\}$，其中，$a_1$ 代表数据块中所有属性值的个数，a_2 代表相同值的记录数，a_3 代表不同值的记录数，a_4 代表相同值连续出现的平均长度，a_5 代表空值的数目，a_6 代表数据是否有序，a_7 代表数据类型。

定义 6 先验知识。先验知识是指在建立贝叶斯分类模型之前，对数据所在列族选取一定比例的训练样本，通过对这些样本进行分类统计得到的数据，这些数据是基于朴素贝叶斯理论的选择策略中分类器的分类基础。

定义 7 评估层。评估层是指在建立贝叶斯分类模型之前，根据压缩结果对先验知识进行调整的一个过程，在评估层中依据数据块压缩后的整体压缩率对训练样本作重新选择，主要是对训练样本所占的比例、单个训练样本的位置和训练样本的大小进行调整。

定义 8 轻量级压缩算法和重量级压缩算法[5]。轻量级压缩算法是指可以直接访问压缩后数据的压缩算法，如字典编码（dictionary encoding）、游程编码（run-length encoding）和位向量编码（bit-vector encoding）；重量级压缩算法是指不可以直接访问压缩后数据的压缩算法，如 HBase 数据库自带的压缩算法 Gzip 编码和 LZO 编码。

4 基于 HBase 数据分类的压缩策略选择方法研究

图 1 所示为本文提出的基于 HBase 数据分类的压缩策略选择方法（CSSM-BCHD, compression strategies selection method based on classification of hbase data）的过程，主要包括：冷热数据分类、压缩策略选择、数据压缩过程和底层存储。其中，数据压缩过程由相邻块推荐策略压缩、基于贝叶斯方法的压缩和自学习压缩组成。

4.1 冷热数据分类处理

以往的研究很少涉及冷热数据的分类压缩，鉴于冷热数据对于压缩策略的要求不同，本文综合了压缩率、数据访问频度和查询性能等几个方面，对不同访问级别的数据进行策略推荐，构建了在 HBase 中基于冷热数据分类的压缩策略选择方法。首先，对待压缩数据进行分类，划分为若干访问级别；然后，对不同访问级别的数据根据事先制定的算法采用合适的策略进行压缩。

目前，需要存储的数据有冷数据和热数据之分。冷数据是指很少访问或不再访问的一类数据，而热数据是需要频繁访问的一类数据[2]。大数据的存储需要区分冷热数据，这对节省存储空间和成本是很有意义的。在所要存储的数据中，将近 80%的数据都是冷数据[2]，针对这样的一类数据，需要找到合适的压缩策略来进行压缩存储。如何确定数据的冷热性分类并进行级别划分，需要一个合理的模型对数据进行分类，决策树是一个理想的分类模型。决策树的优势在于构造过程不需

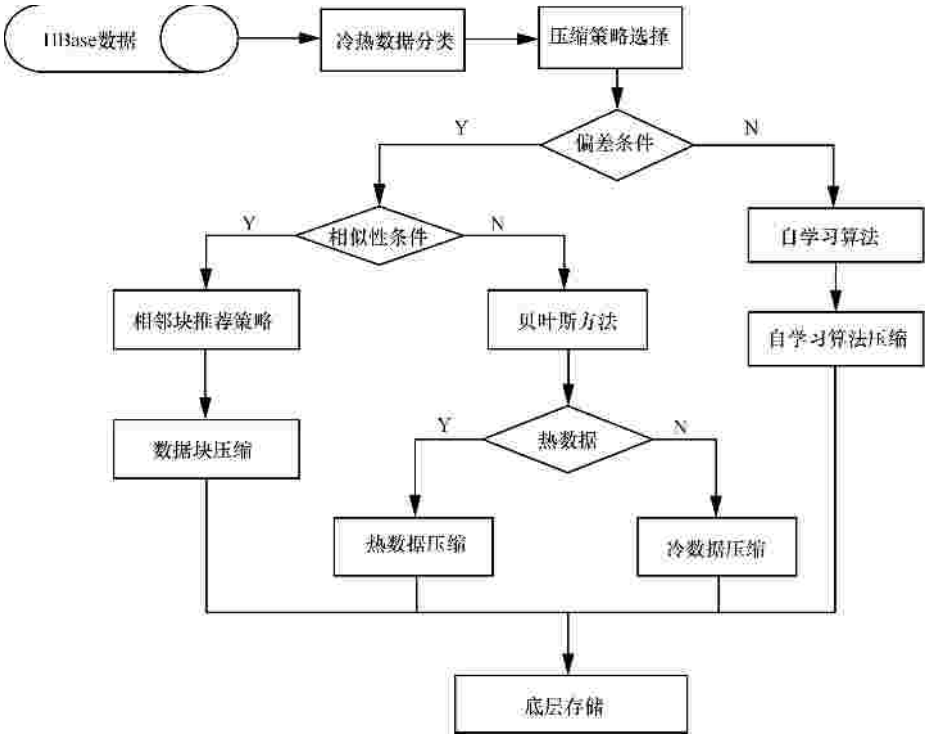

图 1　CSSM-BCHD 流程

要任何领域知识或参数设置，因此在实际应用中，对于探测式的知识发现，决策树更加适用[14, 15]。首先，建立一个模型，描述预定的数据类集或概念集；其次，利用建立的模型将要存储的数据进行分类。由于 HBase 数据库是基于列存储的，并且每一列的数据都是存储在文件系统的同一个文件中[11]。可利用这样的特点对 HDFS 文件系统中文件的访问次数进行统计，以此来作为划分数据访问级别的依据。

构造冷热数据分类决策树模型过程包括以下 4 部分。

1) 在树的每个节点使用信息增益（information gain）度量选择测试属性，选择具有最高信息增益的属性作为当前节点的测试属性。信息增益的定义为

$$Gain(A) = I(s_1, \mathrm{L}, s_m) - E(A) \tag{1}$$

2) 找出一个包含 s 个数据样本的集合 S，s 的具体值应由待分类的数据集的大小决定，与数据集的大小成正比。假定类标号属性具有 m 个不同值，定义 m 个不同类 C_i，设 s_i 是类 C_i 中的样本数。对一个给定的样本分类计算所需的期望信息，由式(2)给出。

$$I(s_1, s_2, \mathrm{L}, s_m) = -\sum_{i=1}^{m} P_i \mathrm{lb}(P_i) \tag{2}$$

其中，P_i 是任意样本属于 C_i 的概率，并用 $\frac{s_i}{s}$ 估计。设属性 A 具有 v 个不同值$\{a_1, a_2, \mathrm{L}, a_v\}$。在这里，属性 A 可以是数据冷热性和热数据访问频度。利用属性 A 将 S 划分为 v 个子集$\{s_1, s_2, \mathrm{L}, s_v\}$。其中，$s_j$ 包含 S 中这样一些样本，它们在 A 上具有值 a_j。如果 A 选作测试属性，则这些子集对应由包含集合 S 的节点生长出来的分支。设 S_{ij} 是子集 s_j 中类 c_i 的样本数。

3) A 划分子集的熵或期望信息由式(3)给出。

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \mathrm{L} + s_{mj}}{s} I(s_{1j}, \mathrm{L}, s_{mj}) \tag{3}$$

其中，$\frac{(s_{1j} + \mathrm{L} + s_{mj})}{s}$ 充当第 j 个子集的权，并且等于子集中的样本个数除以 S 中的样本总数。熵值越小，子集划分的纯度越高。对于给定的子集 s_j，有

$$I(s_{1j}, s_{2j}, \mathrm{L}, s_{mj}) = -\sum_{i=1}^{m} P_{ij} \mathrm{lb}(P_{ij}) \tag{4}$$

其中，$P_{ij} = \frac{s_{ij}}{|s_j|}$ 是 s_j 样本属于类 C_i 的概率。

4) 计算每个属性的信息增益，具有最高信息增益的属性选作给定集合 S 的测试属性。创建一个节点，并以该属性标记，对属性的每个值创建分支，

并据此划分样本。本文中的决策树分类模型是将数据分为冷热数据两大类，在此基础上再对热数据进行级别划分。利用构造的决策树将热数据分为热点数据、次热点数据、活跃数据、不活跃数据，而冷数据则对应沉默数据。进行冷热数据分类的决策树由决策节点、分支和叶子组成。决策树中最上面的节点为根节点，每个分支是一个新的决策节点，或者是树的叶子[14]。在这里，每个决策节点代表数据的冷热性或热数据访问频度。

构造冷热数据分类决策树模型（CDTM-BCHDC, construction of decision tree model based on cold and hot data classification）的算法流程描述如下。

CDTM-BCHDC 算法

输入：样本集合 S，属性集合 A={数据冷热属性，热数据 *frequency*}

输出：ID3 决策树模型

Step1 如果所有种类的属性都处理完毕，返回；否则执行 Step 2。

Step2 计算出信息增益最大的属性 a，把该属性作为一个节点。如果仅凭属性 a 就可以对样本集合 S 进行分类，则返回；否则执行 Step 3。

Step3 对属性 a 的每个可能的取值 v，执行以下操作：

1) 将所有属性 a 的值是 v 的样本作为 S 的一个子集 S_v；

2) 生成属性集合 $AT=A-\{a\}$；

3) 以样本集合 S_v 和属性集合 AT 为输入，递归执行 CDTM-BCHDC 算法。

基于以上算法得出的决策树中，根节点代表划分冷热数据的属性，根节点的一个子节点是对热数据进行划分的访问频度属性，余下的叶节点是划分之后的分类结果。

4.2 压缩策略选择

本文提出的压缩策略选择方法是对基于相邻区和基于统计列的策略选择方法的融合。首先对基于统计列的策略选择方法中基于朴素贝叶斯分类的选择方法进行改进，这种方法本质上是对待压缩数据特征信息进行统计继而进行策略选择，选择的准确性很大程度上依赖于训练样本的选取。在该方法中增加了评估层这个过程来及时对先验知识进行调整，评估层是确保数据压缩率的一个重要步骤，其通过对先验知识的调整来保证基于朴素贝叶斯分类的选择方法的准确性。接着，提出了根据数据冷热性和特征信息进行压缩策略选择的方法，该方法根据不同情况，选用基于相邻参照区的压缩策略选择方法或基于统计列的压缩策略选择方法。其中，基于相邻参照区的压缩策略选择方法是对相邻区的数据特征进行统计，并通过学习参照信息与当前区之间的相似性和差异性进行策略推荐。最终实现数据的压缩策略选择。

4.2.1 改进的基于朴素贝叶斯理论的压缩策略选择方法

在提出改进的分类方法前，本文对获取待分类数据样本的抽样方法进行重新定义：分层抽样数学模型是将数据总体中各个数据块个体按某种特征分成若干个互不重叠的几部分，每一部分称为一个层，在各层中按层在总体中所占比例进行简单随机抽样或系统抽样[16]。针对冷热数据的分布特性，该方法能有效克服样本不能准确反映总体特征的不足。

基于朴素贝叶斯理论的策略选择方法包括以下 3 层。

1) 评估层。根据压缩的结果调整先验知识，对训练样本作重新选择。通常设置一个变量 p，p 为一个数据块压缩后的整体压缩率，再设置一个变量 t，t 是一个阈值（t 的初始值为对同一块数据使用 Gzip 和 LZO 的压缩率的平均值，若前一次数据块的压缩率大于 t，将其值赋予 t），当 $p<t$ 时，应对训练样本进行调整。训练样本的调整方法为：对数据所在列训练样本所占的比例进行调整（每次在原来的基础上增加 $\frac{1}{20}$，即 $n=n\left(1+\frac{1}{20}\right)$，对单个训练样本的位置进行调整（训练样本的选择是均匀的），对训练样本的大小进行调整（单个训练样本规模应在原来的基础上缩减，$m=m\left(1+\frac{1}{20}\right)$）。

2) 分类器训练层。根据特征属性对训练样本进行统计，将各训练样本划分到对应的算法类别中（选择相应的压缩策略），统计各个压缩策略出现的概率及各个特征分量在各个策略下的条件概率。

3) 应用层。对实际的数据块选择条件概率最大的压缩策略。朴素贝叶斯方法的具体实现如下。

首先要构建一个分类集合，集合里的每一个类别绑定一个认为是最优的算法，本文构建的分类集合有 6 个类别，绑定的压缩编码方式有游程编码、字典编码、位向量编码、LZO 编码、Gzip 编码和无

压缩编码。其次采用朴素贝叶斯分类器来划分数据所属的压缩算法类别，朴素贝叶斯分类器的分类目标是在给定特征值和有限的目标集合的前提下，选择最可能的目标值，也就是概率最大的目标。

根据贝叶斯公式有

$$p(v_i \mid a_1, a_2, \mathrm{L}, a_i, \mathrm{L}, a_n) = \frac{p(a_1, a_2, \mathrm{L}, a_i, \mathrm{L}, a_n \mid v_i) p(v_i)}{p(a_1, a_2, \mathrm{L}, a_i, \mathrm{L}, a_n)} \tag{5}$$

将实例 $\boldsymbol{T}$ 分配给类 v_i 的条件就是

$$p(v_i \mid a_1, a_2, \mathrm{L}, a_i, \mathrm{L}, a_n) > p(v_j \mid a_1, a_2, \mathrm{L}, a_i, \mathrm{L}, a_n), \quad 1 \le j \le m, i \ne j \tag{6}$$

只需 $p(a_1, a_2, \mathrm{L}, a_i, \mathrm{L}, a_n \mid v_i) p(v_i)$ 取值达到最大即可，其中 $p(v_i)$ 为任意一个实例被划分为类 v_i 的概率，可以通过训练样本计算出来

$$p(v_i) = \frac{s_i}{s} \tag{7}$$

其中，s_i 为训练样本中被划分为 v_i 的个数，s 为样本的总数。

根据朴素贝叶斯的一个重要特性，给定实例的各个特性之间相互独立，所以

$$p(a_1, a_2, \mathrm{L}, a_i, \mathrm{L}, a_n \mid v_i) = \prod_{j=1}^{n} p(a_j \mid v_i) \tag{8}$$

其中，$p(a_j \mid v_i)$ 可以通过训练样本计算出来。若 a_j 是离散的属性，那么

$$p(a_j \mid v_i) = \frac{s_{ij}}{s_i} \tag{9}$$

如果 a_j 是连续的属性，一般利用分布函数的概率密度函数处理，其概率密度函数为

$$p(a_j \mid v_i) = g(a_j, m_{v_i}, s_{v_i}) = \frac{1}{\sqrt{2\mathrm{p}s^2}} \mathrm{e}^{-\frac{1}{2s^2}(a_j - m_{v_i})^2} \tag{10}$$

其中，m_{v_i} 表示训练样本的均值，而 s_{v_i} 表示训练样本的方差。

综上，利用朴素贝叶斯方法将待分类样本划分到某一类的过程就是计算在待分类样本的特征出现的前提下，出现概率最大的类别的过程

$$V_{\mathrm{NB}} = \arg\max_{v_i \in V} p(v_i) \prod_{j=1}^{n} p(a_j \mid v_i) \tag{11}$$

$p(v_i)$ 和 $p(a_j \mid v_i)$ 都是可以通过训练样本直接计算出来，所以，只要训练样本选择得当，然后根据待分类样本的特征，就可以将待分类样本划分到想要的分类中。

4.2.2　HBase 数据压缩策略选择方法

本文在基于区级压缩模式的策略选择方法[3]和基于朴素贝叶斯分类的选择方法[13]上进行了折中。当数据特征信息满足不同的条件时，用不同的方法进行策略选择，这样既充分利用了 2 种方法在策略选择上的准确性，同时还规避了 2 种方法各自的不足。对于不满足条件的数据，采用自学习方法来进行策略选择，这样能够保证压缩策略选择的有效性。

在利用算法 CDTM-BCHDC 生成的决策树模型对数据进行分类之后，就需要考虑对不同类型的数据采用不同的压缩策略。对于最常访问的热数据，由于经常被访问，不仅要考虑其压缩率，更要考虑解压时间。这里解压时间的重要性远大于压缩率，故针对相应的数据级别和数据特点选择对应的压缩策略。对于轻量级的压缩算法，在和重量级压缩算法（如 Gzip 和 LZO）查询时间接近的情况下，优先选择轻量级压缩算法，因为可以直接在压缩数据上操作，省去了解压所需的内存开销。在轻量级压缩算法中，在合适的条件下，选择字典算法在压缩率和查询时间上都有优势[5]。对于冷数据，则较常采用压缩率较大的压缩算法，如 Gzip 和 LZO。

由文献[11]可知，HFile 是 HBase 使用的底层存储格式，HFile 对应于列族，一个列族可以有多个 HFile，且每个 HRgion 中，同一列族的数据存储在同一个文件——HStore 中，HDFS 中文件是以数据块 HFile 的形式分配到各个节点进行存储的。因此，在选择压缩策略时，首先对整个文件 HStore 的数据进行分类，然后再对其包含的数据块选择合适的压缩策略。在为数据块 HFile 选择压缩策略之前，需要获得数据块的特征信息。首先确定每个数据块的特征向量 $\boldsymbol{T}$，对第一个数据块，通过自学习得到其对应的压缩策略。对其后的每一个数据块，将该压缩策略对应的统计量分量 q_i 与之前数据块的统计量分量作差后取绝对值，若该绝对值小于事先设置的阈值 k，将与各个训练样本的统计量对应的分量的平均值作差，并取绝对值，比较前后 2 个绝对值的大小，若前者较小，则选择前一个数据块的压缩策略作为本数据块的压缩策略；若后者较小，则采用基于数据分类的朴素贝叶斯方法确定该数据块应采用的压缩策略。若该绝对值大于这个阈值 k，则采用自学习方法确定压缩策略。在进行压缩策略选择之前，需要统计 HDFS 文件 HStore 在最近一段时

间t内的访问次数C并计算访问频度$frequency=\frac{C}{t}$及根据访问频度利用决策树模型进行分类。

基于HBase数据分类的压缩策略选择方法（CSSM-BCHD, compression strategies selection method based on classification of hbase data）的算法流程描述如下。

输入：待压缩HDFS数据文件s

输出：压缩是否成功（0：失败，1：成功）

Step1 判断待压缩数据的冷热性分类，若是热数据，则转Step2；否则，转Step6。

Step2 得到文件s包含的数据块集合$A=\{a_1, a_2, a_3, \text{L}, a_n\}$。设置$m_1$=0，压缩算法$m_{\text{temp}}=m_1$，$i$=1。

Step3 如果a_i是第一个数据块则采用自学习方法选择压缩策略，$m_{\text{temp}}=f(T_i)$；否则，与之前数据块的统计量作差并令$z_1=|q_i-q_{i-1}|$，将z_1与设定的阈值做比较，若大于阈值则采用自学习算法，若小于阈值则转Step4。

Step4 取得各个训练样本的统计量对应的分量的平均值q_+，$z_2=|q_i-q_+|$，根据z_2分别选择推荐方法和基于贝叶斯分类的策略选择方法。

Step5 按照m_i的压缩策略对数据块i进行压缩，若当前数据块不是最后一个数据块，i=i+1，转Step3。

Step6 得到文件s包含的数据块集合$A=\{a_1, a_2, a_3, \text{L}, a_n\}$。设置$m_1$=0，压缩算法$m_{\text{temp}}=m_1$，$i$=1。

Step7 如果a_i是第一个数据块则采用自学习方法选择压缩策略，根据数据具体特征选择Gzip或LZO压缩策略；否则，根据前一个数据块的压缩策略，与之前数据块的统计量分量作差后取绝对值，并令$z_1=|q_i-q_{i-1}|$，若z_1与设定的阈值做比较，若小于阈值则采用推荐方法，m_{temp}=m_{i-1}，若大于阈值则采用自学习方法。

Step8 按照m_i的压缩策略对数据块i进行压缩，若当前数据块不是最后一个数据块，i=i+1，转Step7。

由以上的流程分析可知，CSSM-BCHD方法充分利用了基于相邻参照区和基于统计列的选择方法各自的优势，同一类（冷数据或热数据）数据的特征信息相似或相差不大时，选择推荐方法或基于贝叶斯分类的选择方法，算法的时间复杂度为$O(n)$，空间复杂度为$O(n)$，当待压缩的数据块集合中数据块之间的数据分布特征差异较小时，时间消耗较低；数据分布特征差异较大时，时间消耗较高。由用户对具体的数据块选择合适的压缩策略，在将数据进行压缩处理之后，由云端负责压缩后数据的底层存储，相比于未进行压缩处理的数据存储，这样的存储过程仅仅在用户进行策略选择阶段有一定的时间和空间开销，主要是基于统计列的选择方法中分类器训练层里训练样本和统计信息所需的空间开销，以及评估层中重新选择训练样本所需的时间开销。但在云端的具体存储过程中，整体的数据量明显减少了，由文献[17]可知只有原数据量的40%~60%。

5 仿真实验与结果分析

实验环境为5台虚拟机搭建的集群，其中一台配置为Master节点，硬件配置情况为2.50 GHz CPU，2 GB内存，300 GB硬盘；其他4个为非Master节点，硬件配置情况为1.50 GHz CPU，1 GB内存，400 GB硬盘。集群里的所有机器都安装了Ubuntu 14.04系统和Sun Java 1.7，Hadoop的版本为2.6.0，HBase的版本为0.20.6。仿真实验主要分为2部分：1）从时间消耗、压缩率和查询时间3个方面对改进的基于朴素贝叶斯理论的压缩策略选择方法和基于相邻参照区及基于统计列的压缩策略选择方法进行比较；2）对基于冷热数据分类的压缩策略选择方法进行验证分析。

5.1 实验背景

本实验的测试数据集是采用公认的TPC-H标准数据集[18]。TPC系列基准是数据库领域最为广泛接受的基准，TPC-H是一款面向商品零售业的决策支持系统测试基准，它定义了8张表、22个查询，遵循SQL92标准，TPC-H基准的数据库模式遵循第三范式。通过对TPC-H标准数据集中的数据进行压缩存储来实现以上章节所提出的基于冷热数据分类的压缩策略选择方法。

本实验选取了文献[3]中基于相邻参照区（用NGSS表示）和文献[13]中基于统计列（用SICS表示）的压缩策略选择方法，以及文献[19]中的基于动态字典（用DDSCS表示）的区级压缩策略选择方法分别对相同的数据表进行压缩。对于TPC-H数据集中的数据由DBGen生成数据表LINEITEM作为测试数据集，将NGSS、DDSCS和SICS对LINEITEM表进行压缩。基于朴素贝叶斯理论的压缩策略选择方法（用IMSSBC表示）先利用分层抽样方法从LINEITEM表抽取一定数量的数据样本（称为训练样本）作为训练数据集，再根据特征属性对训练样本进行统计，将各训练样本划分到对应的算法类别

中，统计各个压缩策略出现的概率及各个特征分量在各个策略下的条件概率，最后通过朴素贝叶斯分类方法得到训练样本的先验知识，利用先验知识来决定压缩策略的选择。将 IMSSBC 方法与上述 3 种方法进行比较；另外，为了验证基于冷热数据分类的压缩策略选择方法的有效性，将 IMSSBC 方法和本文提出的 CSSM-BCHD 方法进行比较。

5.2　仿真实验与性能分析

5.2.1　实验 1　改进的压缩策略选择方法性能分析

本实验利用 TPC-H 标准数据集实现基于相邻参照区和基于统计列的压缩策略选择方法。为了评估各个选择方法的优劣，使用 TPC-H 的数据生成工具 DBGen 生成数据表 LINEITEM[20]，LINEITEM 表一共有 59 986 052 行数据，这个数据表一共有 16 个属性列。这 16 个属性列在 HDFS 里一共分为 142 个数据块。本实验对选择方法在时间消耗、压缩率和查询时间 3 个方面分别进行了比较，结果如图 2~图 4 所示。对 LINEITEM 表的每列数据压缩之后再解压，最后计算平均压缩和解压时间，图 2(a)对应的是应用 4 种选择方法得到的平均压缩时间，图 2(b)对应平均解压时间。由图 2 可知，改进的压缩策略选择方法相比于其他 3 种选择方法在压缩时间上有较为明显的优势，而解压时间相差不大。

图 3(a)显示了各个属性列使用不同的策略选择方法得到的压缩效果，而图 3(b)是 4 种选择方法的总体压缩效果。由图 3(a)可知，改进的选择方法在各个列中的压缩效果都比另外 3 种要好，而且数据量越大，这种优势越明显。图 3(b)则清楚地显示了 4 种选择方法的压缩率的情况，改进的压缩策略选择方法压缩率最高，经过改进的选择方法压缩后的数据大小为原数据的 21%左右，基于统计列的选择方法压缩率略低于改进的选择方法，压缩后的数据约为原数据的 28%，基于相邻参照区的选择方法压缩率最低，压缩后的数据为原数据的 43%，基于动态字典的区级选择方法介于基于统计列和基于相邻参照区的选择方法之间，为 35%。

由图 4 可知，在列中的不同属性值个数较少的情况下，基于相邻参照区的选择方法的查询时间较少，查询性能较高。改进的选择方法查询时间在列中不同属性值个数处于 5~10 和 25~40 时查询时间较少，查询性能较高，而基于统计列的查询性能则介于这两者之间，基于动态字典的区级选择方法查询时间要优于基于相邻参照区的选择方法，但在不同属性值个数处于 10~30 之间时查询时间要高于改进的选择方法。

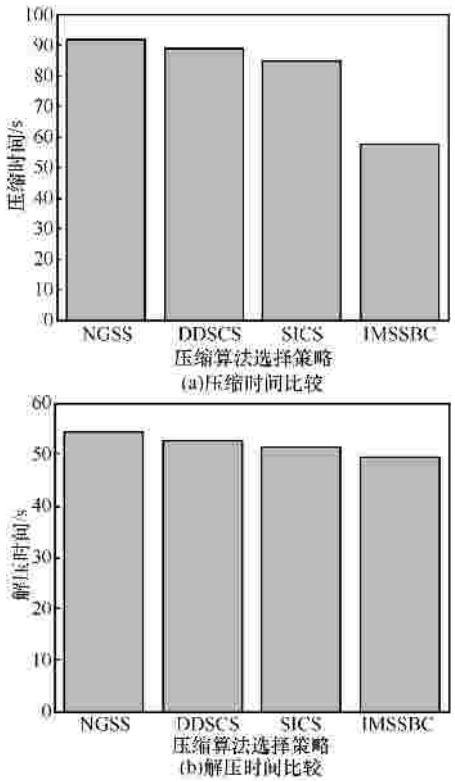

图 2　压缩解压时间比较

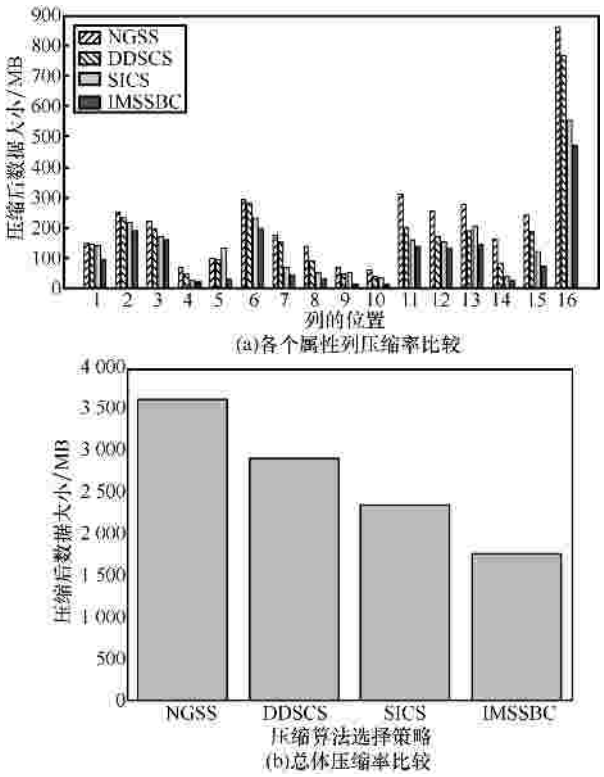

图 3　实验 1 压缩率比较

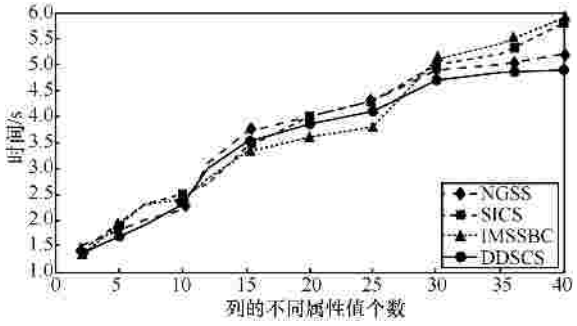

图 4 查询时间比较

5.2.2 实验 2 基于冷热数据分类的压缩策略选择方法性能分析

在对基于冷热数据分类的压缩策略选择方法性能进行分析之前，通过给出基于冷热数据分类的压缩策略选择方法仿真实例，验证基于冷热数据分类的选择方法的有效性。这里还是以数据生成工具 DBGen 生成的数据表 LINEITEM 为例，在对每个列族数据进行压缩存储之前，首先要对列族对应的 HStore 文件进行访问频度的统计，进而根据 4.1 节中提到的 CDTM-BCHDC 算法构造冷热数据分类决策树模型，利用决策树模型得到它的冷热性分类。然后再将 HStore 文件划分为多个数据块，对每个数据块根据特征信息和数据访问级别采用对应的策略进行压缩。

数据表 LINEITEM 各个列族对应的 HStore 文件的访问级别可以通过最近的访问频度得到，列族对应的 HStore 文件访问级别用 Vlevel 表示，冷热性分类结果如表 1 所示。

表 1 LINEITEM 列族对应 HStore 文件访问级别

列	Vlevel	列	Vlevel
1	沉默数据	9	沉默数据
2	不活跃数据	10	沉默数据
3	沉默数据	11	活泼数据
4	沉默数据	12	沉默数据
5	沉默数据	13	沉默数据
6	沉默数据	14	沉默数据
7	沉默数据	15	沉默数据
8	次热点数据	16	沉默数据

将表 1 中各个列族对应的 HStore 文件划分为多个数据块，HStore 文件划分的数据块个数用 N 表示，划分结果如表 2 所示。

表 2 HStore 文件划分的数据块

列	N	列	N
1	9	9	3
2	8	10	3
3	7	11	11
4	3	12	11
5	4	13	11
6	9	14	13
7	6	15	6
8	6	16	26

对每个 HStore 文件的数据块根据特征信息和数据访问级别采用适当的策略，从压缩率和查询时间 2 个方面对基于冷热数据分类的改进压缩策略选择方法的性能进行分析。该实验是在数据表 LINEITEM 上进行数据压缩，对比未进行冷热数据分类的选择方法可得到图 5 和图 6 所示的结果。

图 5(a)和图 5(b)分别从各个列的压缩率和整体的压缩率对数据分类前后的选择方法进行了比较。由图 5(a)可以看出，对于沉默数据的压缩率，基于冷热数据分类的选择方法明显好于未进行分类的选择方法；而对于热数据的压缩率，基于冷热数据分类的选择方法则不如未进行分类的选择方法，这主要是因为基于冷热数据分类的选择方法在选择压缩策略时不仅要考虑压缩率，还要兼顾解压时间和查询性能等因素。从图 5(b)可知，从整体上看，对于数据表 LINEITEM，基于冷热数据分类选择方法的压缩率要好于未进行分类的选择方法，这是因为数据表中 81.7%的数据都是冷数据，对这些数据采用基于冷热数据分类的选择方法的压缩率是比较好的。

在查询时间方面，仿真实验 2 将 IMSSBC 方法和 CSSM-BCHD 方法进行了对比分析，这样做的主要原因是：IMSSBC 方法是建立在未对数据进行冷热性划分及访问级别分类基础上的，而 CSSM-BCHD 方法则是对数据进行了分类和级别划分之后再压缩的，由于 NGSS 和 SICS 与 IMSSBC 方法在查询性能上差异不明显，此处只采用 IMSSBC 方法和 CSSM-BCHD 方法进行分析比较。图 6 显示了 2 种方法在数据表中各个列上查询时间的结果。由于实验 2 侧重于 IMSSBC 与 CSSM-BCHD 算法的比较，这 2 个算法在列值较小时能发现查询效果的差异，故选择的列值范围是 1~16，不同于实验 1 的

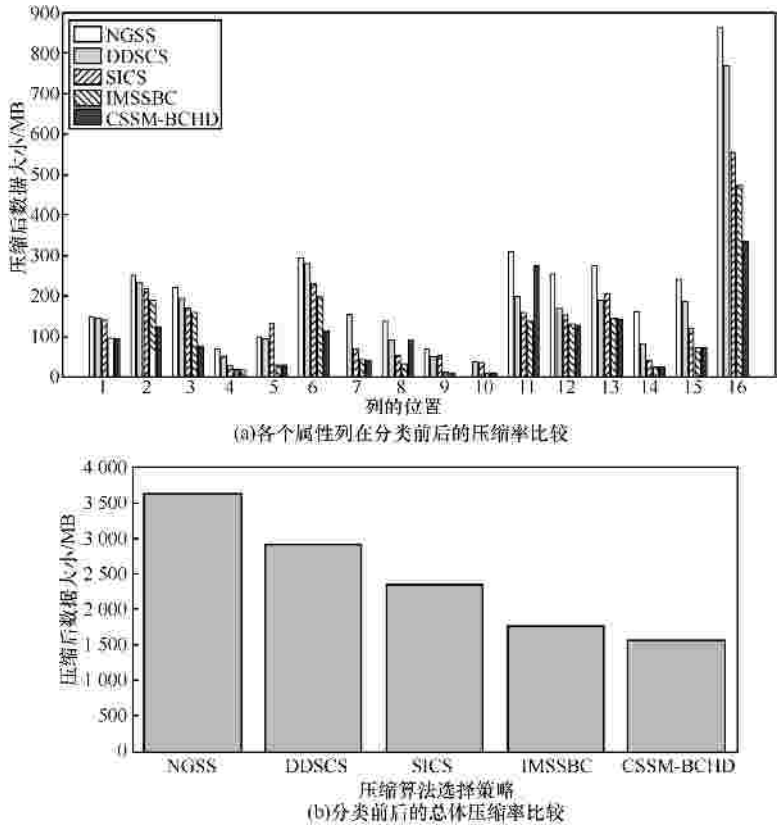

图 5　实验 2 压缩率比较

查询时间的横坐标。从图 6 中可以看出，对于沉默数据，采用基于冷热数据分类的选择策略压缩后的数据查询性能并不比未进行分类的选择策略好，但它们之间的差距不明显；而对于热数据，基于冷热数据分类的选择策略压缩后的数据查询性能要好于未进行分类的选择策略，对经常访问的数据，查询性能的好坏直接影响到整体的查询效率，对于规模比较大的数据，这种效果更明显。

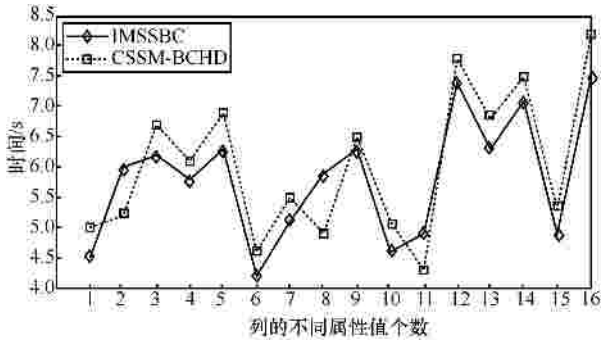

图 6　查询时间比较

6　结束语

HBase 数据存储是大数据处理的关键部分，数据压缩是 HBase 数据存储的重要研究内容，合理的压缩策略选择方法决定了 HBase 数据压缩存储效果。本文针对数据的访问频度不同，提出了数据冷热性的概念；对基于相邻参照区和统计列的压缩策略选择方法进行分析比较，提出了改进的压缩策略选择方法；在此基础上，提出了基于冷热数据分类的策略选择方法。实验表明，本文基于冷热数据分类的选择方法在提高了 HBase 的存储效率的同时，也提高了数据的整体访问性能。本文的下一步工作是：1）将基于冷热数据分类的策略选择方法压缩过的数据采取某种策略迁移到合适的存储设备上；2）在冷热数据分类模型——决策树模型中，考虑加入冷数据访问频度属性作为决策节点。

参考文献：

[1] 程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述[J].软件学报, 2014, 25(9): 1889-1908.
CHENG X Q,JIN X L,WANG Y Z, et al. Survey on big data system and analytic technology[J].Journal of Software,2014,25(9):1889-1908.

[2] 郭嘉凯. 如何存储“冷数据”[J].软件和信息服务, 2013, 23(10): 58-59.
GUO J K. How to store “cloud data”[J]. Software and Information Service, 2013, 23(10): 58-59.

[3] **王振玺，乐嘉锦，王梅，等. 列存储数据区级压缩模式与压缩策略选择方法**[J]. **计算机学报**, 2010, 33(8): 1524-1530.
WANG Z X,LE J J,WANG M, et al. Sector-based compression and compression strategy selection method for column stores[J].Chinese Journal of Computers,2010,33(8):1524-1530.

[4] TRONDHEIM N,STONEBRAKER M,ABADI D J, et al. C-store-A column-oriented DBMS[C] //The 31st VLDB Conference. Trondheim, Norway, c2005: 553-564.

[5] YAN K, XIE M Y ,ZHU H. Fixed-length string compression for direct operations in column-oriented databases[C] //2013 Ninth International Conference on Natural Computation (ICNC). Shenyang,China, c2013: 1171 - 1176.

[6] TOMMY S, SANJAY M. On compressing data in wireless sensor networks for energy efficiency and real time delivery[J].Distributed and Parallel Databases,2013,31(2):151-182.

[7] MEHLA U S, DASGUPTA K S. Hamming distance based reordering and columnwise bit stuffing with difference vector: a etter scheme for test data compression with run length based codes[C] //VLSID 23rd International Conference on VLSI Design. India, Bangalore, c2010: 33-38.

[8] **丘建平，张广艳，舒继武**. DMStone:**一个分级存储系统性能测试工具**[J].**软件学报**, 2012, 23(4): 987 - 995.
QIU J P, ZHANG G Y, SHU J W. DMStone: a tool for evaluating hierarchical storage management systems[J]. Journal of Software, 2012, 23(4):987-995.

[9] LEVANDOSKI J J, LARSON P A, STOICA R. Identifying hot and cold data in main-memory databases[C]//2013 IEEE 29th International Conference on Data Engineering (ICDE).Australia, Brisbane, c2013: 26-37.

[10] GAO H B, WANG D F. LSD2H: a novel storage method of linked sensor data based on HBase[C]//2014 10th International Conference on Semantics, Knowledge and Grids (SKG). Beijing, China, c2014: 116-119.

[11] **朱敏，程佳，柏文阳. 一种基于** HBase **的** RDF **数据存储模型**[J].**计算机研究与发展**, 2013, 50(Suppl.)：23-31.
ZHU M ,CHENG J, BAI W Y. A storage model for RDF data based on HBase[J].Journal of Computer Research and Development, 2013, 50(Suppl.)：23-31.

[12] **葛微，罗圣美，周文辉，等**. HiBase:**一种基于分层式索引的高效** HBase **查询技术与系统**[C]//2014 **中国大数据技术大会. 中国，北京**, c2014.
GE W, LUO S M,ZHOU W H, et al. HiBase: a hierarchical ndexing mechanism and system for efficient HBase query[C]//Big Data Technology Conference 2014.China, Beijing, c2014.

[13] IDREOSS.Self-organizing tuple reconstruction in column-stores[C]// Proceedings of the SIGMOD. Providence, Rhode Island, USA, c2009: 297-308.

[14] **宁正元，王李近. 统计与决策常用算法及其实现**[M]. **北京: 清华大学出版社**, 2009: 260-345.
NING Z Y, WANG L J. Statistical and decision algorithm and its realization[M]. Beijing: Tsinghua University Press, 2009: 260-345.

[15] **李光，王亚东，苏小红. 隐私保持的决策树分类挖掘**[J].**电子学报**, 2010, 38(1): 204-212.
LI G, WANG Y D, SU X H. Privacy preserving data mining on decision tree[J].Acta Electronica Sinica,2010, 38(1): 204-212.

[16] **崔颖安，李雪，王志晓，等. 在线社交媒体数据抽样方法的比较研究**[J].**计算机学报**, 2014, 37(8): 1859-1876.
CUI Y A, LI X,WANG Z X, et al. A comparison on methodologies of sampling online social media[J].Chinese Journal of Computer, 2014, 37(8): 1859-1876.

[17] WEI Y, YONG W F.A Cost-effective and reliable cloud storage[C] //2014 IEEE International Conference on Cloud Computing. Anchorage, AK, c2014: 938 - 939.

[18] TPC[EB/OL]. http://www.tpc.org/tpch/ 2011.

[19] **龙磚涛.列存储数据仓库中压缩技术的研究与实现**[D]. **上海: 东华大学**, 2013.
LONG B T. Research and implementation of compression technology in column-oriented data warehouse[D]. Shanghai: Donghua University, 2013.

[20] NYAMAGWA M,LIU J, UEHARA T. Cloud foren: a novel framework for digital forensics in cloud computing[J].Journal of Harbin Institute of Technology,2014,21(6):39-45.

作者简介：

王海艳（1974-），女，江苏东台人，南京邮电大学教授，主要研究方向为服务计算、可信计算、大数据应用与云计算技术、隐私保护技术。

伏彩航（1990-），男，江苏连云港人，南京邮电大学硕士生，主要研究方向为大数据应用与云计算技术。