

基于云计算的多源综合信息 大容量快速存储算法

刘琰[†], 刘文龙, 许磊, 刘君宇

(北京经济管理职业学院, 北京 100102)

摘要:云计算为IT行业带来新的技术革命,在网络体系内具有更广泛的互操作性。基于云计算的多源综合信息大容量快速存储算法,以异构方案设定读写流程,通过加权处理均衡大容量信息负载,在给定的数据存储逻辑结构的前提下,对照字节长度划分信息存储单元,基于云计算提出分类存储条件,按照相似性判断准则,快速存储多源综合信息。实验结果表明:以视频和音频信息作为测试对象,按照多种属性调取样本信息,在本文算法应用下可以完成快速存储,且在转化后能够保证数据的完整度,具有应用效果。

关键词:云计算;多源综合信息;大容量;快速存储

中图分类号:P628.4

文献标识码:A

A Large Capacity Fast Storage Algorithm for Multi-Source Comprehensive Information Based on Cloud Computing

LIU Yan[†], LIU Wenlong, XU Lei, LIU Junyu

(Beijing Institute of Economic And Management, Beijing 100102, China)

Abstract: Cloud computing has brought a new technological revolution to the IT industry. It has more extensive interoperability in the network system. Based on cloud computing multi-source comprehensive information high-capacity fast storage algorithm, this study set the read/write process in a heterogeneous scheme, balance the large capacity information load through weighted processing, divide the information storage unit according to the byte length, propose classified storage conditions based on cloud computing, and quickly store multi-source comprehensive information according to the similarity judgment criteria on the premise of a given data storage logical structure. The results show that taking video and audio information as the test object, sample information can be retrieved according to various attributes, which can be quickly stored under the application of the algorithm in this paper, and the integrity of the data can be guaranteed after conversion, with application effect.

Key words: cloud computing; multisource comprehensive information; high-capacity; fast storage

信息储存技术经过了两次发展浪潮:第一阶段以处理为中心,将处理器作为核心发展动力,促使计算机行业快速普及;第二阶段以传输为中心,将网络作为发展动力,在互联网的普及中快速获取信息数据。在两次浪潮推动下,加快了信息数字化的进程,大量的信息数据转化为数字形式,在网络中

呈现爆炸式增长,进而推动数字技术的革新和发展。伴随网络的普及和应用,互联网的访问量和数据量快速增长,因此对存储系统的要求越来越高,需要将大量的信息数据进行快速存储。为了保证网络中大容量信息的完整性和存储的迅速性,各种存储技术和产品应运而生,文献[1]中提出了并源

收稿日期:2022-11-15

基金项目:北京经济管理职业学院 2022 年校级科研项目(22BYA01)

作者简介:刘琰(1985—),女,河南清丰人,硕士,助理研究员,研究方向:大数据分析,创新创业教育,云计算等。

[†] 通信联系人, E-mail: yuelumieno@163.com

数据集群的存储技术,直接通过数据库建立关系框架,直接在访问过程中完成数据信息缓存,但在访问量持续增加的过程中,会引发许多问题。如数据库需要在不断拆库和拆表的循环中对访问数据进行扩充,导致数据库的架构发生转变,并在成本上受到限制。针对互联网多源信息的产生和融合,在信息存储和调取中存在问题,如负载过重和响应时间长等,为保证不同用户的响应需求,需要设计一个新的信息存储方式,实现大容量数据的快速存储,在高并发状态下减少网络的时延。本文以此为研究基础,应用云计算技术设计存储算法,对多源综合信息进行存储,按照四大基本存储结构(顺序存储、链接存储、索引存储和散列存储),在存储体系结构层次不同层级引入网络和存储抽象,依据数据计算和存储需求持续扩张,即区分多个分类存储类型,对应多源综合信息的存储单元,整个分类存储过程与多源综合信息类型分为四个部分:

- (1)根据业务类型:对应高并发数据。
 - (2)根据虚拟节点的映射关系:按照不同的编号形式,对能够产生数据的物理机器进行标记,对应数据 IP 资源。
 - (3)根据不同的字节长度:对应字节流和字符流以及混合流。
 - (4)根据的特征复杂度:对应类似数据的分组。
- 对应所属的存储单元,为实现大容量数据的快速存储提供理论支持。

1 加权处理下均衡大容量信息负载

在网络数据激增的情况下,互联网的访问量呈现高并发状态,若想将所有的数据信息进行存储,需要在现有的网络中应用快速存储数据库,能够针对高并发场景,实现多种类型数据的存放^[2]。以 REDIS 数据异构方案,对应互联网中数据主库,按照同步读写形式,在互联网运行中形成映射框架,当用户在互联网中发出 HTTP 请求后,直接将读作操作投入数据主库中,则可以直接通过异构实现访问。将指令下发给客户端,实现传输路径中的同步存储,使得互联网平台中数据,能够在对应关系下保持数据的同步性。

而对多源综合信息数据的存储,需要根据业务类型完成区分,在设计数据服务组组长时,保证数据主库与异构方案之间的转换关系,能够在高并发场景下,对多种类型的数据完成划分和存储^[3]。在数据主库与传输通路中,将抽取语言技术作为主要指令标准,不考虑中间件的影响,在互联网产生数据信息时,直接在通路中发出存储指令即可。在面

对高并发数据压力下,为保证数据的负载均衡,以加权一致性的分配方式,对多源综合信息数据集群进行整合,利用虚拟节点的映射关系,在固定间隔时间内,将虚拟节点按需分配到服务器中:

$$Z = X \times \beta(\lg X)$$
 (1)

$$C(V) = \frac{B_V}{SUM \times Z}$$
 (2)

式中:虚拟节点总数为 Z 。分类参数为 β 。物理机器总数为 X 。物理机器的资源数为 C 。数据编号为 V ,则实际物理机资源数为 B_V ^[4]。以 $\log X$ 对数关系获取节点信息,即存在如下关系:

$$SUM = \sum X_V$$
 (3)

在虚拟节点中对应数据 IP 资源,按照不同的编号形式,将能够产生数据的物理机器进行标记,以每次获取的最大负载率和最小负载率差值,作为下次启动的负载均衡值^[5]。以 N 、 M 两组服务器产生的负载作为均衡条件,设定其负载率为 A_N 、 A_M ,服务器的负载均值分别为 A_{AVG} ,最大负载与最小负载分别为 A_{max} 、 A_{min} 。

在加权一致性处理原则下,对服务器的数据负载进行调节,具体见表 1^[6]。

表 1 加权一致性处理原则下负载均衡顺序

序号	内容	解释
1	$while(A_N - A_M)$	a. 统计 REDIS 方案中的服务器虚拟节点; b. 降序排列负载大小
2.	$\{(A_{AVG} - A_M), (A_N - A_{AVG})\}$	a. 服务器自身负载与均值负载比较; b. 转移高负载虚拟节点
3	$A_N = A \rightarrow REFRESH(N)$	更新服务器 N 的负载率
4	$A_M = A \rightarrow REFRESH(M)$	更新服务器 M 的负载率
5	$A_N = \max\{A(V) V = 1, 2, \dots, X\}$	确定服务器服务类型
6	$A_M = \min\{A(V) V = 1, 2, \dots, X\}$	a. 编辑服务器虚拟节点编号; b. 重新分配数据负载

如表中内容所示,在互联网中对应出数据传输的服务器,以虚拟节点对照数据情况。根据虚拟节点实现映射,使数据的资源数与节点保持线性增加,在发生较大负载时,能够转移到下一个节点中,使得数据资源达到负载均衡。在此基础上,将分配后的数据负载机械能分类,根据不同的数据字节长度,对应所属的存储单元。

2 对照字节长度划分信息存储单元

多源综合数据的存储是一项较大的挑战,主要是数据的体量较大且数据类型繁多,需要在快速处理中完成存储。只有具备规则有序的数据结构,才

能为后续的数据分析提供便利条件,因此在数据存储中需要保证信息,按照其特定的类型完成分类^[7]。在大数据研究领域中,存储的数据信息包含字节流、字符流、混合流三个类型,此次根据主流形式设计多源综合信息的存储结构,划分信息存储单元。数据存储单元即单独目录形式的存储结构,按照四个层次进行分类,压缩前的存储单元逻辑结构。

将存储单元分为数据文件、数据页、数据行、数据字段四个层次,分别为 DF 、 DP 、 DR 、 $DFID$ 。在不同的存储单元中存在复杂的数据结构,需要在分类前设置虚拟搜索指令,在存储时能够针对数据类型匹配相应的索引^[8]。其中在每一个存储单元中含有一个或多个 DF ,按照数据列进行划分,每个 DF 中会含有多个 DP ,而在 DP 中又存在有对应的 DR ,并分裂成多个 $DFID$ 。以标准的类型对应数据字节长度,其数据类型如下:

$$\mathbf{D}_1 = [1\text{byte}] \quad (4)$$

$$\mathbf{D}_2 = \begin{bmatrix} 1\text{byte}_1 & 1\text{byte}_2 \\ 2\text{byte}_1 & 2\text{byte}_2 \end{bmatrix} \quad (5)$$

$$\mathbf{D}_4 = \begin{bmatrix} 1\text{byte}_1 & 1\text{byte}_2 & 1\text{byte}_3 & 1\text{byte}_4 \\ 2\text{byte}_1 & 2\text{byte}_2 & 2\text{byte}_3 & 2\text{byte}_4 \\ 3\text{byte}_1 & 3\text{byte}_2 & 3\text{byte}_3 & 3\text{byte}_4 \\ 4\text{byte}_1 & 4\text{byte}_2 & 4\text{byte}_3 & 4\text{byte}_4 \end{bmatrix} \quad (6)$$

式中: $\mathbf{D}_1$ 表示单字节存储结构。 $\mathbf{D}_2$ 为双字节整型结构。 $\mathbf{D}_4$ 为四字节存储结构。以字节长度定义数据类型,当长度为单字节时,则范围为 $\mathbf{D}_1 = [0, 64)$,当长度为双字节时,范围表示为 $\mathbf{D}_2 = [64, 8192)$,长度为四个字节时,范围取值为 $\mathbf{D}_4 = [4M, 1G)$ ^[9]。

在特定的存储空间中,对存储单元的划分需要确定信息元,通过数据页的最大长度计算字段偏移需要的行量^[10]。以独立字段为逻辑结构设定其在行数据存储中为不连续状态,对综合信息中的资源进行统计,则该空间的总消耗量为:

$$H = (\log 2(K) + \log 2(J)) \times K \quad (7)$$

式中:字段存储的空间消耗量为 H 。信息元为 K 。数据页的最大长度为 J 。在不同的数列中,相同的数据字段可以组织成一个数据页,而相同的数据字段也可以方便数据的压缩,对划分的存储空间释放更多介质空间。通过不同的字节长度对应多源综合信息的存储单元,并利用云计算方式实现数据快速存储。

3 基于云计算快速存储多源综合信息

云计算自身带有信息处理设施,通过互联网向

用户提供服务,并对服务器执行弹性计划,提供不同用户所需的资源和数据。当用户浏览互联网服务器时,在高并发状态下,会产生海量数据,对其完成快速存储,需要根据上文中提出的对应空间,对相似的数据实现合并聚类,条件如下:

$$L_{I=0}^E Q_I = W \quad (8)$$

$$Q_I \cap Q_U = \Phi, I \neq U \quad (9)$$

式中: W 表示一组数据样本。 L 表示样本的相似度。 $Q = \{Q_1, Q_2, \dots, Q_E\}$ 作为数据的聚类簇, Q_I 、 Q_U 是子集,其中 $I, U = (1, 2, \dots, E)$ 。每一个类均带有特征描述, E 为聚类的个数。

将互联网中传输的综合数据存储,认为是多个对象的聚类过程,随机选择 E 个初始对象,通过其与中心距离的长度,作为相似性的判断准则:

$$R = \sum_{I=1}^E \sum_{T \in Q_I} |T - Y_I| \quad (10)$$

式中: R 为所有传输数据的均方差之和。 T 表示多维对象区间 Q_I 中的一个聚类簇节点, Y_I 为聚类 Q_I 的均值。在距离最近的数据类中,可以将其形成一个类型的合并,能够将在一个聚类簇的结构数据,存储在用一个空间中。

对数据的分类要具有恰当性和准确度,通过云计算平台中的同时计算模式,将聚类后的数据特征进行整合,以边缘化描述数据的特征复杂度,计算不同数据的聚类耗时:

$$\begin{aligned} Y_P &= \frac{\alpha_1 xy}{gh} \\ Y_T &= \alpha_2 [x(h-1) + y(g-1)] \\ Y_T + Y_P &= Y_Y \end{aligned} \quad (11)$$

式中: Y_P 、 Y_T 表示所有数据的分类和聚类过程所需的时间。 Y_Y 代表总耗时。 α_1 与为常数,表示聚类和分类的负载量。 α_2 x 、 y 分别表示数据字段的页数和行数。 g 、 h 表示相邻的不同字节数量。当相邻的数据字节量为等同时,可以对公式进行统一化处理:

$$\begin{aligned} Y_{\epsilon P} &= \frac{\alpha_1 x^2}{g^2} \\ Y_{\epsilon T} &= 2\alpha_2 x(g-1) \\ Y_{\epsilon T} + Y_{\epsilon P} &= Y_{\epsilon Y} \end{aligned} \quad (12)$$

此时在相同的数据页数和行数下,以云计算确定多源综合信息的聚类和分类耗时,将 $Y_{\epsilon T}$ 、 $Y_{\epsilon P}$ 、 $Y_{\epsilon Y}$ 定义为线性函数,实现数据的统一化处理。在完成相类似数据的分组后,对多源综合信息放置在独立空间中,完成不同类型数据的存储。至此,通过数据的负载均衡。数据字段的划分以及云计算的数据聚类,实现快速存储算法设计。

4 实验测试分析

上文通过云计算方式完成了快速存储算法设计,为验证其能够有效进行大容量数据快速存储,采用对比测试完成论证。选择 AUV 算法、蚂蚁算法为对照组,分别与本文算法进行比较,以多源信息接入设备作为测试对象,在不同的数据输入中进行信息存储和分类。整个测试按照两个阶段进行:第一阶段比较不同算法的存储时效,设计多个分类存储类型,在不同的预处理环境下,对传输的信息数据进行整理并储存,比较各组算法的性能;第二阶段比较数据信息的存储状态,即在完成全部类型信息的分类的存储后,对照给定的数据源,验证其是否按照类型完成精准存储。通过设计的两种测试环节,不仅可以验证存储算法的效率,也可以在后续应用时保证多源综合信息的存储完整度。

4.1 多源综合信息数据准备

为保证数据来源的真实性,选择一组多源信息接入设备作为测试对象,其中包含视频输入装置、视频转换装置、音频输入装置和音频切换装置,并且在各装置终端连接有传感器。调取其不同属性的信息数据,具体情况见表 2。

表 2 多源信息接入设备数据统计(组、G)

属性类型	视频输入装置	视频转换装置	音频输入装置	音频切换装置
A1	1400	1.4	2600	2.6
A2	1600	1.6	2800	2.8
A3	1800	1.8	3000	3.0
B1	2500	2.5	3800	3.8
B2	3000	3.0	4000	4.0
B3	3500	3.5	4200	4.2
C1	5000	5.0	5500	5.5
C2	6000	6.0	6500	6.5
C3	7000	7.0	7500	7.5

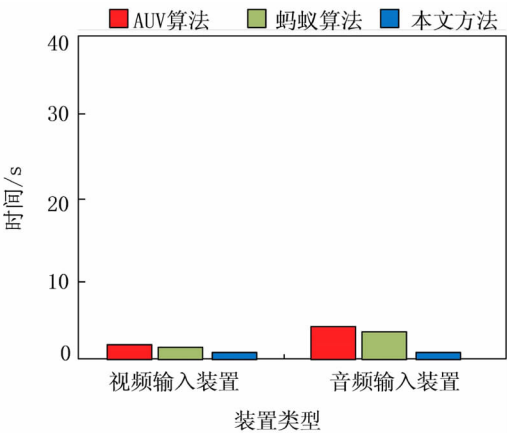
根据表中内容,在选择接入设备中,每一种样本属性个体,均存在有大量的数据信息,符合大容量信息存储测试要求,各属性数据大小为 $A_{总} < B_{总} < C_{总}$ 。并且在视频和音频输入装置中,随着其接收数据的增加,其转换数据的含量也增加,只有保证输入数据在短时间内完成快速存储,才能够在转换装置中完成数据类型的切换,将信息传达和显示。

音频和视频信息在切换后,会呈现为压缩状态,以压缩包的形式分类至不同属性中,当用户发出检索和调取指令后,再将压缩包完成解压和传输,因此需要保证在存储过程中音频和视频信息的完整度。下文通过选择的 3 种算法进行测试,将调

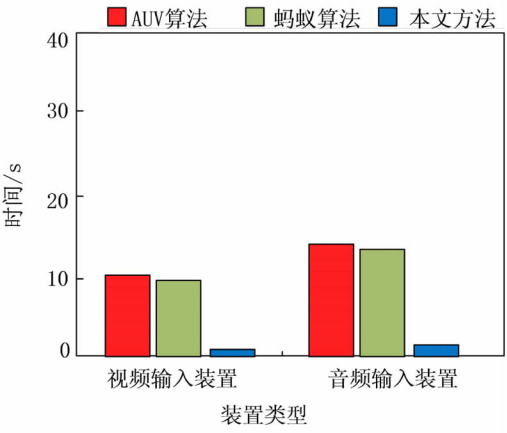
取的样本数据上传至 MATLAB 平台,分别连接存储算法,验证不同方法的有效性。

4.2 不同分类条件下信息存储时效分析

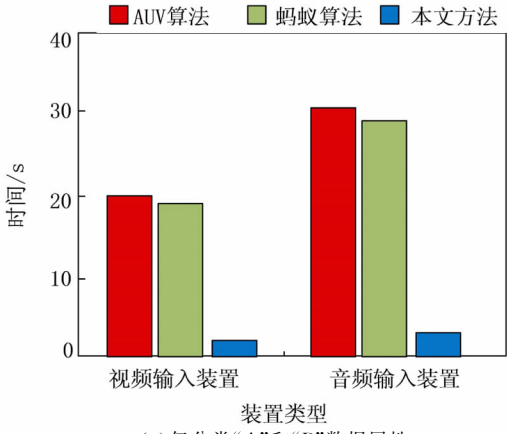
从抽取的样本数据可知,在不同属性的大容量数据中,每个属性的数据均含有三组类型,以是否分类存储作为测试条件,在不同的分类标准下完成存储算法测试。分别按照不分类数据属性存储、仅分类“A”数据属性存储、仅分类“A”和“B”数据属性存储以及全部属性均分类存储进行测试,具体结果如图 1 所示。



(a) 不分类数据属性



(b) 仅分类“A”数据属性



(c) 仅分类“A”和“B”数据属性

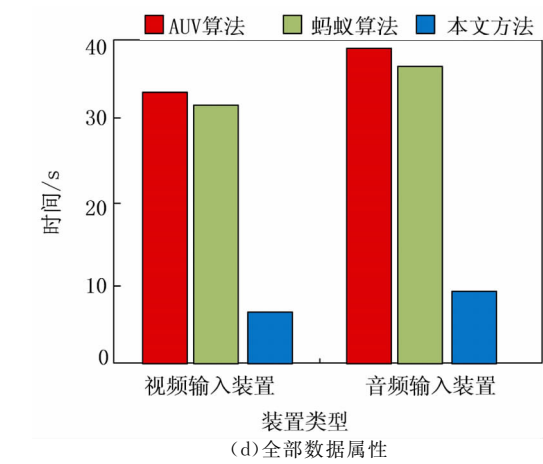


图 1 存储时效对比结果

根据图中内容,在不同存储要求下,各算法呈现的时效不同,在不分类的前提下,各组算法耗时较为接近,均可以控制在 5s 之内。在给定存储条件时,本文算法的耗时改变量较小,仍可以保证在 10s 内完成信息存储,而其他两组算法的存储耗时增加量较大,尤其在全部分类条件下,两组传统算法的耗时超过了 30s。综合结果表明:本文方法能够对大容量数据进行合理分类,具有快速存储信息的效果。

4.3 不同算法应用下信息存储完整度分析

在完成第一阶段测试后,对每组算法下存储的数据完整度进行验证,即使各算法能够在一定时间内实现数据存储,但若数据存储过程中丢失了部分数据,在后续应用时会产生负面影响。以各组属性的总转换量为测试基础,测试结果见图 2。

根据图中内容,只有本文算法能够保证各属性数据存储的完整度,其他两组传统方法均发生了少量的数据丢失情况。综合全文内容可知:本文方法在具备快速存储的能力同时,能够准确地实现数据分类,保证存储信息的完整度,具有应用价值。

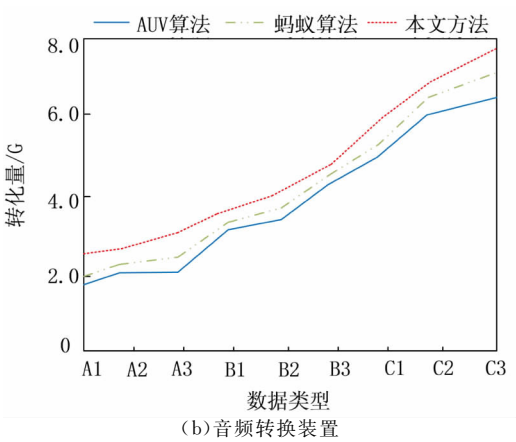
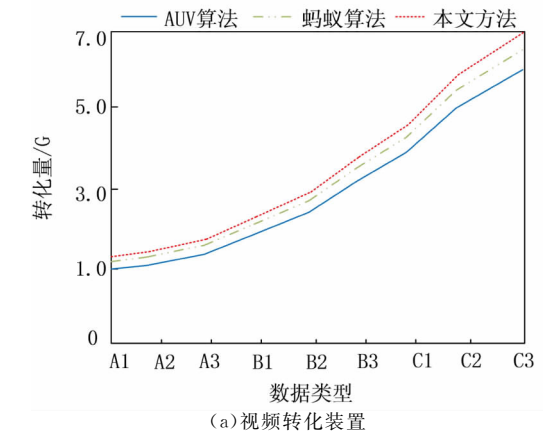


图 2 存储完整度对比结果

5 结 论

在海量暴增的网络数据中,只有快速且完整地实现数据存储,才能够保证后期的数据挖掘有效性,为保证多源网络中海量数据的快速存储,本文以云计算方式为基础,设计了新的多源综合信息存储方法,并通过实验验证了其有效性,可以在不同的数据类型中实现快速存储。但由于时间有限,在设计中未考虑数据存储过程的安全性,即在网络攻击中仍可以进行快速存储,后续研究中会进一步分析,为数据的存储提供理论支持。

参考文献

[1] 刘灯明,荆俊峰,刘凯,等. 基于改进蚁群算法的云计算资源分配策略研究[J]. 电子技术应用,2022,48(5):104-109.

[2] 张金泉,徐寿伟,李信诚,等. 基于正交自适应鲸鱼优化的云计算任务调度[J]. 计算机应用,2022,42(5):1516-1523.

[3] 谭印,苏雯洁. 基于深度哈希算法的云计算虚拟机迁移模型[J]. 计算机仿真,2022,39(2):366-370.

[4] 秦宝堃,郑洪英,肖迪. 基于预测误差双重编码的大容量密文域可逆信息隐藏算法[J]. 计算机应用研究,2022,39(1):245-248.

[5] 黄紫如,柴洪洲,向民志,等. 考虑信息滞后的 AUV 因子图多源信息融合定位算法[J]. 中国惯性技术学报,2021,29(5):625-631.

[6] 鲁琴. 基于云计算技术的舰船信息系统数据多批处理调度方法[J]. 舰船科学技术,2020,42(6):172-174.

[7] 史振华. 基于改进的鲸鱼算法在云计算任务调度中的研究[J]. 科技通报,2021,37(2):67-71.

[8] 刘宇,惠鸿飞,路永乐,等. 基于遗传算法优化神经网络的多源信息融合室内定位方法[J]. 中国惯性技术学报,2020,28(1):67-73.

[9] 杨晓梅. 多源信息数据的并行优化抽样聚类 K-means 算法[J]. 计算机仿真,2020,37(7):305-308+332.

[10] 何登平,张为易,黄浩. 基于多源信息聚类 and IRC-RBM 的混合推荐算法[J]. 计算机工程与科学,2020,42(6):1089-1095.