文章编号:1671-5896(2020)04-0457-10

量子计算在增量式大数据并行挖掘中的应用

李晓峰[1], 王妍玮[2], 李 东[3]

(1. 黑龙江外国语学院 信息工程系, 哈尔滨 150025; 2. 普度大学 机械工程系, 印第安纳州 西拉法叶市 IN47906;
3. 哈尔滨工业大学 计算机科学与技术学院, 哈尔滨 150001)

摘要: 针对传统大数据并行挖掘方法是一次性对所有数据进行挖掘, 导致挖掘时间较长, 挖掘精度较低等问题, 采用量子计算对增量式大数据并行挖掘方法进行优化设计。首先, 按照数据挖掘的基本流程搭建并行数据挖掘模型; 然后分别通过定义量子比特、量子搜索算法、量子神经网络处理以及量子映射变换4个步骤, 实现增量式数据的预处理, 利用矩阵向量相乘分解得到过滤权重组合, 通过该组合实现预处理结果的并行协同过滤; 最后通过量子模糊聚类得出增量式大数据并行挖掘结果。实验结果表明, 应用量子计算的增量式大数据并行挖掘方法的平均召回率为97.25%, 并行挖掘时间在2.1~3.2 s的范围内浮动, 准确率超过95%, 且该方法的收敛性最好, 寻优能力强。

关键词: 量子计算; 增量式; 大数据; 并行挖掘

中图分类号: TP391 **文献标识码**: A

Application of Quantum Computing in Incremental Parallel Mining of Large Data

LI Xiaofeng[1], WANG Yanwei[2], LI Dong[3]

(1. Department of Information Engineering, Heilongjiang International University, Harbin 150025, China;
2. Department of Mechanical Engineering, Purdue University, West Lafayette, Indianan IN47906, US;
3. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: In view of the problem that the traditional big data parallel mining method always mines all the data at one time, resulting in a long mining time and low mining accuracy, the quantum computing is adopted to optimize the incremental big data parallel mining method. Firstly, the parallel data mining model is built according to the basic process of data mining. Then on the mining model respectively by defining a quantum bit, quantum search algorithm, quantum neural network processing and mapping transformation, the incremental data preprocessing, filtering weights are obtained by decomposition of matrix-vector multiplication, preprocessing results by using the combination of parallel collaborative filtering. Finally, by quantum fuzzy clustering, large incremental data parallel mining results are obtained. The experimental results show that the average recall rate of the incremental big data parallel mining method using quantum computing is 97.25%, the parallel mining time is within the range of 2.1~3.2 s, and the accuracy rate is always above 95%. And this method has the best convergence and strong optimization ability.

Key words: quantum computing; incremental; big data; parallel mining

收稿日期: 2019-10-17

基金项目: 国家自然科学基金资助项目(61803117); 教育部科技发展中心产学研创新基金资助项目(2018A01002); 国家科技部创新方法专项基金资助项目(2017IM010500)

作者简介: 李晓峰(1978—), 男, 哈尔滨人, 黑龙江外国语学院教授, 博士, 主要从事人工智能、机器学习和智慧医疗等研究, (Tel)86-451-88121567(E-mail)lixiaofeng@hiu.net.cn。

0 引 言

大数据本身是一个比较抽象的概念，主要指需要新处理模式才能具有更强的决策力、洞察力和流程优化能力的改良、高增长率和多样化的信息资产。由于在互联网大数据中能描绘出具体的数据特征，因此对其进行数据挖掘具有较大的研究价值。在通常情况下大数据具有数据体量大、数据类型繁多、处理速度快以及价值密度低的特点，因此在对其挖掘过程中需要经过几个基本的挖掘步骤，才能得到最终结果。通常在数据挖掘过程中，首先进行数据采集，然后通过数据清理、集成、归约以及变换等步骤，完成数据的预处理过程，最终经过特征提取与处理、数据建模与转换，输出数据挖掘的最终评估结果。由于其数据信息不是静态的，而是以增量式形式逐渐增加的，因此为了提高数据挖掘效率，近些年人们提出了并行数据挖掘技术，通过将顺序执行的计算任务分解成多个可以同时执行的子任务，并且其可以同时并行执行，最终完成整个计算任务。

量子算法是一种遵循量子力学规律调控量子信息单元进行计算的新型计算模式，张焕国等[1]主要分析了量子问题、量子线路及量子算法复杂性的有关问题，对量子环境下的算法设计和问题求解具有指导意义；孙晓明[2]围绕着量子算法、计算复杂性、程序理论、电路和密码学，对近些年来量子力学理论展开研究与总结，并展望了未来量子算法的发展方向；许精明等[3]根据量子算法具有计算灵活，结果准确的优点，将量子算法应用于网络节点扩展的酉变换矩阵的研究中；金贻荣等[4]主要研究了超导量子计算在抑制电荷、磁通和准粒子等几种主要噪声来源方面做出探索；张兰等[5]主要对量子粒子群优化算法进行了研究；周海鹏等[6]主要研究了自适应混沌量子粒子群算法及其在 WSN(Wireless Sensor Network)覆盖优化中的应用；张惠珍等[7]在传统量子算法的基础上，利用量子蚁群算法分析了多车次同时送取货物车辆路径问题。

诸多专家学者对增量式大数据并行挖掘方法进行了研究，并取得了一定的研究成果。文献[8]提出一种基于 MapReduce 的并行频繁模式增量挖掘算法。由于传统的频繁模式挖掘算法是以“批处理”方式执行的，即一次性对所有数据进行挖掘，无法满足不断增长的大数据挖掘的需要，因此引入 MapReduce，将传统频繁模式增量挖掘算法 CanTree 向 MapReduce 计算模型进行了迁移，实现了并行的频繁模式增量挖掘。文献[9]提出一种大数据环境下关联规则并行分层挖掘算法，将整个数据库 D 随机分割成若干个非重叠区域，并行挖掘出局部频繁项集，利用先验性质连接局部频繁项集得到全局候选项集，再次扫描 D 统计出每个候选项集的实际支持度，以确定全局频繁项集，以此实现大数据并行挖掘。文献[10]提出了时空大数据背景下并行数据处理分析挖掘方法，以其作为研究基础，重点从时空大数据的存储管理、时空分析和领域挖掘 3 个角度对并行化挖掘方法进行了设计。

经过研究发现，上述并行式数据挖掘方法无法考虑到大数据增量变化的问题，在增量式大数据中召回率低，不能全面且准确地挖掘大数据中的数据信息。为解决上述问题，笔者在传统并行数据挖掘的基础上，引入量子算法，提出基于量子算法的增量式大数据并行挖掘方法。将量子算法应用于大数据挖掘方法的优势在于打破了图灵机模式的计算限制，提高指数效率和计算性能，以达到保证增量式大数据并行挖掘的基础上实现大数据的精准挖掘。

1 构建增量式大数据并行挖掘模型

按照大数据并行挖掘的基本流程搭建相应的并行数据挖掘模型，并在模型中引入量子算法，其模型的搭建结果 MapReduce 如图 1 所示。图 1 中并行数据挖掘模型实现的基本过程一般有：数据准备、开采、评估与表示等。通过该模型，大数据并行挖掘的相关程序将被自动的分布到一个由普通机器组成的超大机群上并发执行，而模型中的 Map 和 Reduce 分布为该模型中的两大基本操作，分别表示数据挖掘中的映射过程和归约过程。Map 任务在执行过程中需要将一组数据一对一的映射到另外一组数据中，而 Reduce 利用映射规则与归约规则实现数据聚类挖掘[11]。两个步骤程序通过把大数据的操作分发给网络上的每个节点，实现数据的并行式挖掘，每个节点以周期性的运行方式，将计算得到的结果和状态信息返回到主节点中。

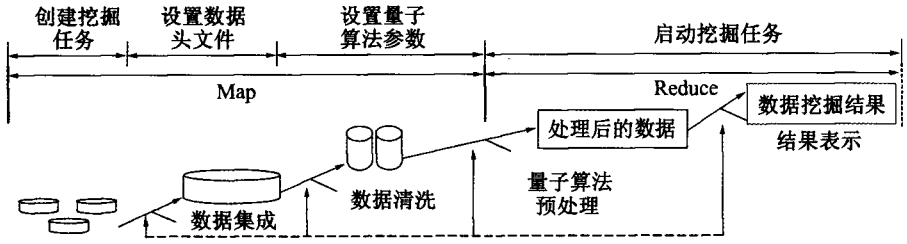

图 1　数据挖掘模型

Fig. 1　Data mining model

2　量子算法数据预处理

量子计算是综合利用量子力学原理和计算机科学相关知识进行计算的一种新的计算模式。在现阶段利用量子机制进行信息处理已成为突破经典计算极限的一条重要的探索途径，量子计算建立在与传统比特概念类似的量子比特上，主要利用量子系统的叠加性、纠缠性和相干性等实现量子的并行计算。

通过量子算法进行数据的预处理，将计算模型上的数据进行移植，使用 3 个 MapReduce 任务实现并行挖掘预处理。预处理基本原理如图 2 所示。

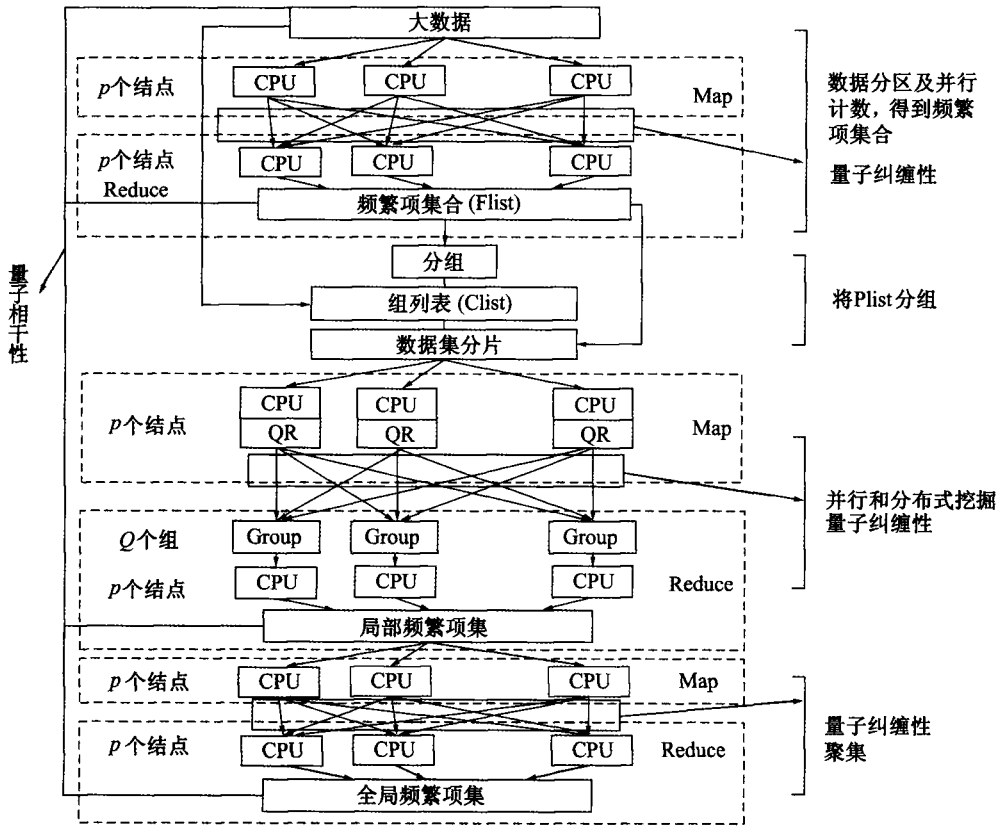

图 2　量子算法基本原理图

Fig. 2　Basic principle of quantum algorithm

在图 2 中，不同结点按照其性质划分至不同组别，结点所对应的组别不是唯一的，属于一对多关系，体现了量子纠缠性。对大数据进行数据分区与并行计数，获取大数据频繁项集合，进而将该集合分为局部频繁项与全局频繁项，体现了数据之间的关联性，说明其具有量子相干性。在对大数据处理过程中，其频繁项可以处在不同量子态的叠加态上，因此体现了量子叠加性。

按照图 2 中的量子算法原理，使用 HDFS(Hadoop Distributed File System)将大数据进行自动分配，在第 1 个任务执行中对所有数据项进行并行计算，得到所有频繁项的集合，记为 FList。将分组过的频繁项

集合记为组列表 GList。将包含组列表的集合进行加载启动，并对执行结点进行数据分片处理，输出每组关联事物结果。最后将所有的局部频繁项合并在一起得到全局频繁项集，即为大数据挖掘的预处理结果。按照量子算法的基本计算原理[12]，对预处理步骤进行具体分析。

2.1 量子比特的数据信息转换

量子算法在应用处理过程中只能针对量子比特数据进行计算，量子比特也是量子算法中存储信息的位移格式。因此需要将大数据中的数据信息转换成量子比特下的单位格式，用于描述量子线路的状态信息，因此首先对量子比特进行定义。根据量子力学的基本原理，将量子比特的状态定为0态或1态。此外，量子比特的状态也可以落在0态和1态之外，可以是任意线性叠加的状态[13]。1个n位的量子存储器可以处于2^n个基态的相干叠加态$|\varphi\rangle$中，由此便可以同时存储2^n个不同数字，对量子寄存器进行一次操作就相当于对传统计算机的2^n次操作实现并行挖掘。其中相干叠加态$|\varphi\rangle$可表示为

$$|\varphi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{1}$$

其中α和β分别表示一对复数，即量子态的概率幅值[14]。α和β需要满足

$$|\alpha|^2 + |\beta|^2 = 1 \tag{2}$$

的条件。因此可以将量子态的概率幅值表示为$[\alpha,\beta]^{\mathrm{T}}$。将量子比特用球面坐标的形式表示，复数$\alpha$和$\beta$用$\cos\frac{\theta}{2}$和$e^{i\varphi}\sin\frac{\theta}{2}$代替，由此可得

$$|\varphi\rangle = \cos\frac{\theta}{2}|0\rangle + e^{i\varphi}\sin\frac{\theta}{2}|1\rangle \tag{3}$$

其中θ与φ定义了三维单位球面上的一点P，如图3所示。

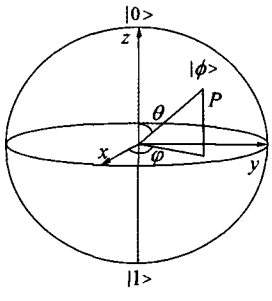

图3 量子比特坐标示意图

Fig. 3 Schematic diagram of qubit coordinates

将量子比特定义在球面上使单个量子比特状态可视化，方便进行量子计算与量子信息识别，以此实现量子比特。

2.2 量子搜索算法

在量子算法实际应用过程中难以做到使搜索概率达到100%，且量子相位旋转角的误差会导致搜索概率下降，因此要对量子搜索算法进行改进。设量子搜索算法初始状态为$\Psi_m\phi$，经过m迭代后得到

$$\Psi_m\phi = |\theta\Psi(k_x^m, l_x^m)\phi \tag{4}$$

其中k_x^m与l_x^m分别表示初始状态与理想状态经过多次迭代后的变化幅度，θ为量子相位旋转角。

搜索过程中的相位旋转角θ是影响量子搜索算法性能的主要因素，因此利用

$$\theta' = 2\arcsin\left[\sin\frac{\pi}{4J+2}/\sin\beta\right] \tag{5}$$

对其进行改进，其中J为量子理想态数。

设大数据并行挖掘中的待挖掘数据的标记态为$|s_1\rangle, |s_2\rangle, \cdots, |s_M\rangle$，而非标记态用$|t_1\rangle, |t_2\rangle, \cdots, |t_{N-M}\rangle$，将对应系数代入可得

$$|o\rangle = \frac{1}{\sqrt{N}}\sum_{i=0}^{N-1}|i\rangle = \frac{\cos\theta'}{\sqrt{N-M}}\sum_{i=1}^{N-M}|t_i\rangle + \frac{\sin\theta'}{\sqrt{M}}\sum_{i=1}^{M}|s_i\rangle \tag{6}$$

将两个相移算子I_s与I_0进行推广描述，可将量子搜索[15]结果描述为

$$\boldsymbol{G} = -HI_0H^+I_s = \begin{pmatrix} (1-e^{i\alpha})\cos^2\theta' - 1 & e^{i\alpha}(1-e^{i\alpha})\sin\theta'\cos\theta' \\ (1-e^{i\alpha})\sin\theta'\cos\theta' & e^{i\alpha}(1-e^{i\alpha})\sin^2\theta' - 1 \end{pmatrix} \tag{7}$$

其中H表示变换算子，通过量子搜索算法得出数据挖掘的初始数据。将得到的搜索结果记为$\boldsymbol{G} = (g_1, g_2, \cdots, g_n)^{\mathrm{T}}$且在搜索过程中保证搜索的成功概率恒等于1。

2.3　量子神经网络处理

在上述量子搜索中经常要同时搜索多个模式，会在一定程度上降低量子搜索精度，因此在量子搜索算法的相位旋转角进行改进的基础上，利用量子神经网络解决该问题。

量子神经网络中当模拟人的意识时要处理多模式问题，能有效提升量子搜索精度与速率。量子神经网络结构如图4所示。

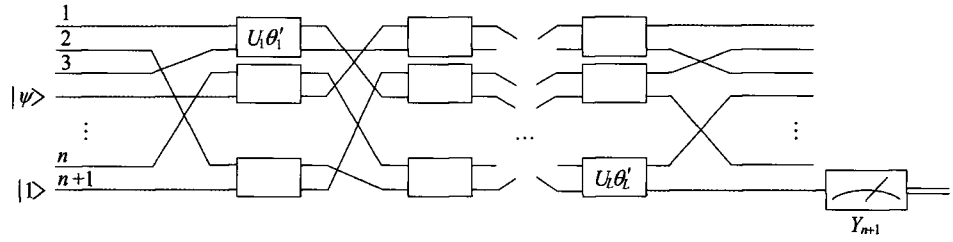

图4　量子神经网络结构

Fig. 4　Structure of quantum neural network

基于上述量子神经网络结构，将得到的数据搜索结果经过量子比特转换得

$$|\boldsymbol{G}\rangle = [\,|g_1\rangle, |g_2\rangle, \cdots, |g_n\rangle\,]^{\mathrm{T}} \tag{8}$$

其中 $\boldsymbol{G}$ 表示量子比特转换方程。且

$$|g_i\rangle = \cos\left(\frac{2\pi}{1+\exp(-g_i)}\right)|\psi\rangle + \sin\left(\frac{2\pi}{1+\exp(-g_i)}\right)|1\rangle \tag{9}$$

其中 g_i 表示量子网络参数。按照量子网络中的数据参数的更新规则，对数据参数进行多模式搜索，得到归一化后的期望输出

$$\boldsymbol{G}(t+1) = \eta\Delta\,\boldsymbol{G}(t) \tag{10}$$

其中 η 为量子神经网络的学习更新速率，$\Delta\,\boldsymbol{G}(t)$ 为更新梯度值[16]。

利用量子神经网络完成数据归一化处理，获取数据更新处理结果。

2.4　量子映射变换

经过量子神经网络更新处理完成的数据即为量子算法预处理的结果数据，然而更新的数据是以量子比特的形式输出，因此需要通过量子映射变换，将数据转换为大数据中的正常形式。将量子映射变换过程表示如下

$$U_m = I - 2\sum_{i=1}^{m}|\alpha_i\rangle\langle\alpha_i| \tag{11}$$

其中 I 为量子映射参数，α_i 为量子基态。根据量子映射变换公式输出增量式大数据预处理结果如下

$$G_{ij}^{\theta} = \alpha_i + \frac{(\alpha_i - \beta_i)\theta_i'}{\pi|\varphi\rangle} \tag{12}$$

其中 $i=0,1,2,\cdots,n$，得出的 G_{ij}^{θ} 全局频繁项集合即为增量式大数据并行挖掘的预处理结果。

3　数据并行协同过滤

由上述得到的增量式大数据并行挖掘的预处理结果为基础，利用数据并行协同过滤实现得到数据并行挖掘结果，其过程为：计算待挖掘数据之间的相关度，得到相似度的计算集合，根据该集合分解大数据的权重矩阵，得到过滤权重组合，利用该组合对增量式大数据并行挖掘的预处理结果进行数据并行过滤，得到增量式大数据并行挖掘的初始结果。首先将所有的大数据的权重构建成为 $x\times y$ 的矩阵[17]，并用向量的形式进行表示，可表示为 $\boldsymbol{R}(u=x,y)=(R_{u1},R_{u2},\cdots,R_{un})$。然后利用相似性计算公式，求出待挖掘数据之间的相关度，计算公式如下

$$S_{\mathrm{sim}}(u,v) = \frac{\sum\limits_{i\in R(u)\cap R(v)}(R_{ui}-\bar{R}_u)(R_{vi}-\bar{R}_v)}{\sqrt{\sum\limits_{i\in R(u)\cap R(v)}(R_{ui}-\bar{R}_u)^2}\sqrt{\sum\limits_{i\in R(u)\cap R(v)}(R_{vi}-\bar{R}_v)^2}} \tag{13}$$

其中 $\bar{R}_u$ 与 $\bar{R}_v$ 分别代表数据集 u 与 v 中所有数据的均值[18]。根据相似度的计算结果集合，将矩阵与向量相乘，进行大数据的权重矩阵分解，其过程如图5所示。

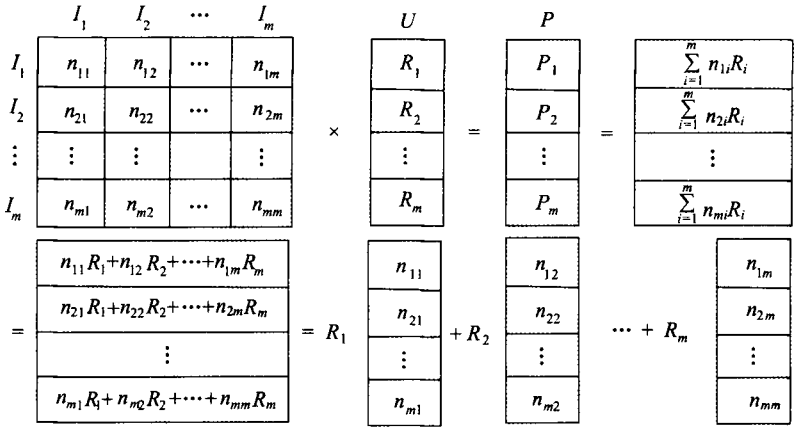

图5 矩阵向量相乘分解图

Fig. 5 Matrix vector multiplication decomposition diagram

将共现矩阵中项目对应的列向量，与数据向量中对应的相似度相乘，即可得到过滤权重组合[19]，最后将各个部分相加即可得到完整的并行协同过滤结果 $\boldsymbol{R}'$，将并行协同过滤的结果作为静态大数据环境下的数据并行挖掘结果。

4 数据增量式聚类挖掘结果

在静态挖掘结果的基础上，利用大数据增量式技术与量子算法中的增量性质，实现对增量式大数据的遍历，将每次遍历得到的挖掘结果进行模糊聚类，便可得到增量式大数据并行挖掘结果。应用的增量式技术是指对新增数据，利用已有结果对其进行处理，即避免了对已有数据的重新处理，又可以对新增的数据进行全面处理，使每次处理的数据量很小，且使用时间更短，大幅度提升聚类效果。

在此过程中会出现策略不适用的问题，为解决该问题，只需修改因数据变化而涉及的规则即可完成数据的增量式处理。

数据增量式模糊聚类将静态的数据挖掘结果用一个代价函数进行迭代，聚类原理如图6所示。

遵循图中的量子模糊聚类原理，静态数据挖掘的结果 $\boldsymbol{R}'$ 分为 N 个集合数据与 k 个类别 C_i，定义静态挖掘结果 $\boldsymbol{R}'$ 中的聚类中心为 P_i[20]，且 i 的取值为[1,k]。根据聚类的类间分离原则，将增量式聚类结果描述为

$$f(\boldsymbol{R}') = \min \sum_{i=1}^{k} \sum_{\boldsymbol{R}' \in C_i} \| \boldsymbol{R}' - P_i \| \cap \max \sum_{i,j=1 \,\&\, i \neq j}^{k} \| P_i - P_j \| \tag{14}$$

其中 P_j 表示 $\boldsymbol{R}'$ 中任意一个数据，对式(11)进行求解，便可得出增量式大数据挖掘聚类结果为

$$F = \frac{\sum_{i=1}^{k} \sum_{\boldsymbol{R}' \in C_i} \| \boldsymbol{R}' - P_i \|}{\sum_{i,j=1 \,\&\, i \neq j}^{k} \| P_i - P_j \|} \tag{15}$$

将聚类结果以编码的形式输出便可得到增量式大数据并行挖掘结果，其中聚类结果的编码电路如图7所示。

第1位比特的态即为单比特逻辑态，第2~6位量子比特处于态$|0\rangle$。在图7中将第6位比特设置控制位，利用该控制位实现对其他比特位的控制操作，完成聚类结果的编码，输出增量式大数据并行挖掘结果。

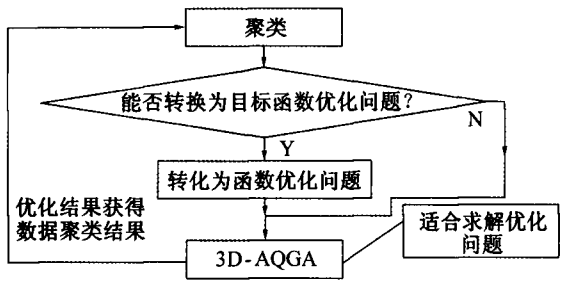

图 6　量子模糊聚类原理图

Fig. 6　Principle diagram of quantum fuzzy clustering

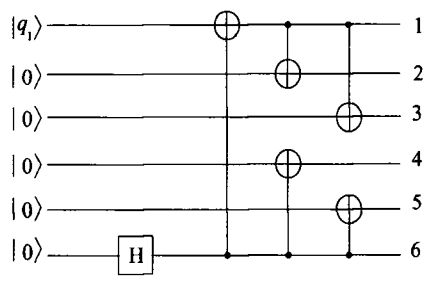

图 7　量子编码电路

Fig. 7　Quantum coding circuit

5　实验结果与分析

5.1　实验环境

由于该实验需要对海量的数据进行处理，因此对实验环境具有较高的要求。计算机处理器为酷睿 i7 处理器，运行存储空间为 8 GByte，内存空间为 64 GByte，并增加外部存储设备为 80 TByte。在符合实验要求的设备上安装实验所需的实验方法和对比方法。其中对比方法为不应用量子算法的传统增量式大数据并行挖掘方法。

5.2　实验数据集

为了检测量子算法的增量式大数据并行挖掘方法的综合性能，进行了实验测试，选取 GroupLens (https://grouplens.org/datasets/movielens/)提供的数据集作为实验样本数据。该大数据中包含了不同大小与类型的多个数据集，挖掘出用户对不同领域的需要具有重要的商业意义。数据集中部分数据描述如表 1 所示。

表 1　数据集部分数据描述

Tab. 1　Data set partial data description

大数据名称	数据数量/GByte	数据维度	标准类数	大数据名称	数据数量/GByte	数据维度	标准类数
Iris	2 379	12	12	Education	1 024	4	2
Wine	5 311	22	11	Entertainment	2 436	12	6
Glass	6 548	16	8	Traffic	1 024	12	4
Ad	3 564	28	7	Tourism	3 256	6	3
Text	1 024	34	17	Shopping	2 324	32	16
Medical	8 462	42	21	Food	1 024	4	4

实验中所用的数据存在的形式为标准格式，可以直接应用于实验中。

5.3　实验指标

由于文献[8-10]方法都是增量式大数据并行挖掘较新研究，也是具有代表性的研究成果，因此选择这 3 种方法进行实验测试，实验中选取的主要评价指标如下。

1）召回率。此次对比实验主要为了检测挖掘方法的性能，因此选择文献[8]方法与笔者方法进行数据挖掘的召回率对比。召回率的计算公式如下

$$R_{\text{recall}} = \frac{T_P}{T_P + F_N} \tag{16}$$

其中 T_P 表示的是实验数据挖掘的有效结果，F_N 为数据挖掘的漏报数据结果。因此在实验过程中通过记录实验的有效结果数据与漏报结果数据便可以得出两种数据挖掘方法的召回率。

2）并行挖掘时间。并行挖掘时间是反应数据挖掘效率高低的重要因素，因此将笔者方法与 3 种方法进行并行挖掘时间对比。

3）并行挖掘准确率。并行挖掘准确率是衡量数据挖掘方法性能的主要因素之一，因此将笔者方法与 3 种方法进行并行挖掘准确率对比。

4) 寻优能力。寻优能力是反应并行挖掘结果好坏的主要因素，将笔者方法与3种方法进行寻优能力对比。在寻优过程中最优迭代次数计算公式如下

$$M = \frac{\varpi}{\log(F)} \tag{17}$$

其中 F 表示增量式大数据挖掘聚类结果，ϖ 表示实验数据集中的所有数据量。将实验数据数量与大数据挖掘聚类结果代入式(17)中，得出最优迭代次数的计算结果 $M=450$。

5.4 实验结果

将实验数据按照分类分别输入到两种挖掘方法中，文献[8]方法与量子算法下的挖掘方法分别按照各自的方法流程对数据进行处理，最终输出的挖掘结果如表2所示。

表2 实验对比数据

Tab. 2 Experimental comparison data

大数据类型	文献[8]方法			笔者方法		
	有效数据/kByte	漏报数据/kByte	召回率/%	有效数据/kByte	漏报数据/kByte	召回率/%
Iris 大数据	1 985	394	83.43	2 297	82	96.55
Wine 大数据	4 321	990	81.35	5 212	99	98.13
Iris 大数据	5 711	837	87.21	6 358	190	97.09

通过表2可以看出，文献[8]方法的总漏报数据量为2 221 kByte，而应用笔者方法的总漏报数据量为371 kByte。经过计算可知，文献[8]方法的平均召回率为83.99%，而笔者方法的平均召回率为97.25%，比文献[8]方法高13.26%。

在以下实验中，选择的大数据类型随机。为了进一步验证笔者方法的优越性，将3种方法作为实验对比方法进行挖掘时间对比，结果如表3所示。

表3 不同方法并行挖掘时间对比

Tab. 3 Comparison of parallel mining time of different methods

实验次数	时间/s			
	文献[8]方法	文献[9]方法	文献[10]方法	笔者方法
10	5.2	19.8	20.9	3.2
20	5.5	20.5	24.5	2.6
30	5.7	19.3	26.4	2.1
40	5.8	17.5	23.4	2.9
50	6.5	19.7	24.6	2.1
60	5.6	18.7	28.9	2.5

分析表3可知，文献[8]方法的并行挖掘时间在5.2~6.5 s的范围内变化；文献[9]在17.5~20.5 s的范围内变动；文献[10]方法在20.9~28.6 s的范围内变化；笔者方法在2.1~3.2 s的范围内变动。综合比较4种研究方法的并行挖掘时间，笔者方法的挖掘时间最短。

在比较不同研究方法召回率与挖掘时间的基础上，测试不同方法的并行挖掘准确率，结果如图8所示。

分析图8可知，文献[8]方法的并行挖掘准确率在60%~70%之间；文献[9]方法在60%左右，是4种方法中挖掘准确率最低的；文献[10]方法在80%~85%之间；笔者方法在95%以上，与其他研究方法相比具有较高的并行挖掘精度。

在上述实验基础上，为了更全面比较不同研究方法综合性能，需要比较不同方法的算法寻优能力，结果如图9所示。

收敛性主要是指在合适的迭代次数下输出大数据挖掘结果。根据式(17)可知，在迭代次数为450时数据挖掘结果最优。分析图9可知，文献[8]方法最快收敛，其次是文献[10]方法，而文献[9]方法在500次迭代中一直未收敛，笔者方法在440左右停止收敛，因此该方法的收敛性更好。综合比较4种方法可知，笔者方法的收敛性最好，寻优能力强，因此可以得到较好的并行挖掘结果，有利于挖掘出用户

对于不同领域的需要，表明该方法具有重要的商业意义与应用价值。

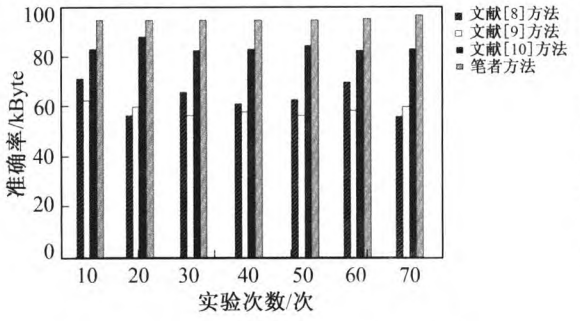

图 8　不同方法并行挖掘准确率对比

Fig. 8　Comparison of parallel mining accuracy of different methods

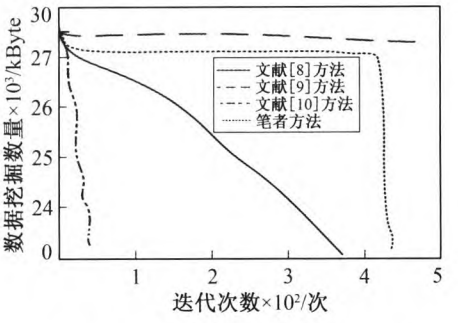

图 9　不同方法寻优能力对比

Fig. 9　Comparison of different methods for optimization

6　结　语

随着大数据时代的到来，数据量正在以惊人的速度增长，从海量数据中挖掘出有效信息，有效组织和利用相关数据都是现阶段需要解决的问题。针对传统增量式大数据并行挖掘方法存在的问题，在增量式大数据并行挖掘方法中应用量子算法，实验结果表明，该方法具有较高的召回率，挖掘时间较短，挖掘准确率较高并具有较好的算法寻优能力。笔者研究方法不仅提高了数据挖掘的召回率，同时也大幅提高了时间效率，为数据挖掘领域的发展及增量式大数据挖掘提供了新思路。

参考文献：

[1]张焕国，毛少武，吴万青，等. 量子计算复杂性理论综述 [J]. 计算机学报，2016，39(12)：2403-2428.
ZHANG Huanguo，MAO Shaowu，WU Wanqing，et al. Overview of Complexity Theory of Quantum Computing [J]. Journal of Computer Science，2016，39(12)：2403-2428.

[2]孙晓明. 量子计算若干前沿问题综述 [J]. 中国科学：信息科学，2016，46(8)：982-1002.
SUN Xiaoming. A Review of Some Frontier Problems in Quantum Computing [J]. Chinese Science：Information Science，2016，46(8)：982-1002.

[3]许精明，阮越. 3-puzzle 量子计算的酉变换矩阵及逻辑线路 [J]. 量子电子学报，2016，33(4)：469-475.
XU Jingming，RUAN Yue. Unitary Transformation Matrix and Logic Circuit of 3-Puzzle Quantum Computation [J]. Journal of Quantum Electronics，2016，33(4)：469-475.

[4]金贻荣，郑东宁. 超导量子计算：长退相干量子比特发展之路 [J]. 科学通报，2017(34)：3935-3946.
JIN Yirong，ZHENG Dongning. Superconducting Quantum Computing：The Way of Long Decoherence Qubit Development [J]. Science Bulletin，2017(34)：3935-3946.

[5]张兰，聂玉峰. 一种融合差分进化的量子粒子群优化算法 [J]. 计算机仿真，2016，33(2)：313-316.
ZHANG Lan，NIE Yufeng. A Quantum Particle Swarm Optimization Algorithm Based on Differential Evolution [J]. Computer Simulation，2016，33(2)：313-316.

[6]周海鹏，高芹，蒋丰千，等. 自适应混沌量子粒子群算法及其在 WSN 覆盖优化中的应用 [J]. 计算机应用，2017，38(4)：1064-1071.
ZHOU Haipeng，GAO Qin，JIANG Fengqian，et al. Adaptive Chaos Quantum Particle Swarm Optimization and Its Application in WSN Coverage Optimization [J]. Computer Application，2017，38(4)：1064-1071.

[7]张惠珍，赵玉苹. 多车次同时送取货物车辆路径问题的量子蚁群算法 [J]. 上海理工大学学报，2017(6)：563-570.
ZHANG Huizhen，ZHAO Yuping. Quantum Ant Colony Algorithm for the Vehicle Routing Problem of Multi-Train Simultaneous Delivery and Pick-up [J]. Journal of Shanghai University of Technology，2017(6)：563-570.

[8]肖文，胡娟，周晓峰. PFPonCanTree：一种基于 MapReduce 的并行频繁模式增量挖掘算法 [J]. 计算机工程与科学，2018，40(1)：15-23.
XIAO Wen，HU Juan，ZHOU Xiaofeng. PFPonCanTree：A Parallel Frequent Pattern Incremental Mining Algorithm based on

MapReduce [J]. Computer Engineering and Science, 2018, 40(1): 15-23.

[9]张忠林, 田苗凤, 刘宗成. 大数据环境下关联规则并行分层挖掘算法研究 [J]. 计算机科学, 2016, 43(1): 286-289.
ZHANG Zhonglin, TIAN Miaofeng, LIU Zongcheng. Research on Parallel Hierarchical Mining Algorithm of Association Rules in Big Data Environment [J]. Computer Science, 2016, 43(1): 286-289.

[10]关雪峰, 曾宇媚. 时空大数据背景下并行数据处理分析挖掘的进展及趋势 [J]. 地理科学进展, 2018, 37(10): 14-27.
GUAN Xuefeng, ZENG Yumei. Progress and Trend of Parallel Data Processing, Analysis and Mining in the Context of Spatiotemporal Big Data [J]. Progress in Geosciences, 2018, 37(10): 14-27.

[11]SONG W, DENG Ze, WANG Lizhe, et al. G-IK-SVD: Parallel IK-SVD on GPUs for Sparse Representation of Spatial Big Data [J]. Journal of Supercomputing, 2017, 73(8): 1-18.

[12]SANG J, GAO Y, BAO B K, et al. Recent Advances in Social Multimedia Big Data Mining and Applications [J]. Multimedia Systems, 2016, 22(1): 1-3.

[13]邵彧. 大数据云存储中的并行优化处理方法仿真 [J]. 计算机仿真, 2016, 33(4): 395-398.
SHAO Yu. Simulation of Parallel Optimization Processing Method in Big Data Cloud Storage [J]. Computer Simulation, 2016, 33(4): 395-398.

[14]谭龙, 张晓琪, 贾立, 等. 一种高效的大数据增量真值发现算法 [J]. 哈尔滨工程大学学报, 2019, 40(4): 805-812.
TAN Long, ZHANG Xiaoqi, JIA Li, et al. An Efficient Incremental Truth Value Discovery Algorithm for Big Data [J]. Journal of Harbin Engineering University, 2019, 40(4): 805-812.

[15]周润物, 李智勇, 陈少淼, 等. 面向大数据处理的并行优化抽样聚类 *K*-means 算法 [J]. 计算机应用, 2016, 36(2): 311-315.
ZHOU Runwu, LI Zhiyong, CHEN Shaomiao, et al. Parallel Optimization Sampling Clustering *K*-Means Algorithm for Big Data Processing [J]. Computer Application, 2016, 36(2): 311-315.

[16]秦静. 一种处理不平衡大数据的并行随机森林算法 [J]. 微电子学与计算机, 2017, 34(4): 22-27.
QIN Jing. A Parallel Random Forest Algorithm for Unbalanced Big Data [J]. Microelectronics and Computer, 2017, 34(4): 22-27.

[17]罗莉. 基于改进 BigFIM 算法的网络信息大数据高频数据项挖掘算法研究 [J]. 激光杂志, 2016(7): 135-140.
LUO Li. Research on High Frequency Data Mining Algorithm of Network Information Big Data Based on Improved BigFIM Algorithm [J]. Laser Journal, 2016(7): 135-140.

[18]李校林, 杜托, 谢勇. 基于 Hadoop 的大数据频繁模式挖掘算法 [J]. 微电子学与计算机, 2018, 35(9): 20-25.
LI Xiaolin, GU Tuo, XIE Yong. Big Data Frequent Pattern Mining Algorithm Based on Hadoop [J]. Microelectronics and Computer, 2018, 35(9): 20-25.

[19]于彦伟, 齐建鹏, 陆云辉, 等. 时空轨迹大数据分布式蜂群模式挖掘算法 [J]. 计算机工程与科学, 2016, 38(2): 255-261.
YU Yanwei, QI Jianpeng, LU Yunhui, et al. Distributed Bee Pattern Mining Algorithm for Spatiotemporal Trajectory Big Data [J]. Computer Engineering and Science, 2016, 38(2): 255-261.

[20]曾俊. 一种基于 Hadoop 架构的并行挖掘算法研究 [J]. 现代电子技术, 2018, 41(1): 117-119.
ZENG Jun. Research on A Parallel Mining Algorithm Based on Hadoop Architecture [J]. Modern Electronic Technology, 2018, 41(1): 117-119.

(责任编辑: 张洁)