基于 Spark 并行架构的多源异构学习行为数据融合模型研究

王崇科，任　刚，魏　勇
（河南工学院 计算机科学与技术学院，河南 新乡 453003）

摘要：当前教育信息化进程中的学习行为数据呈现出明显的多源异构性和大规模性，对数据融合计算提出较大挑战。该文首先设计了一个具有三层结构的学习行为标准数据集，包括元数据层、行为数据层和上下文数据层，为异构数据语义整合提供统一框架；然后基于此数据集，构建了一个基于 Spark 并行架构的多源异构学习行为数据融合模型 Spark-DataFusion。实验结果表明，该模型具有良好的数据规模扩展性和计算节点扩展性，能够有效提升大规模学习行为数据处理效率，对促进教育大数据深度挖掘具有积极意义。

关键词：Spark；学习行为数据；数据融合；并行计算
中图分类号：TP18；TP311.13　　**文献标志码**：A　　**文章编号**：2096-7772-(2025)05-0014-05

0　引言

随着教育信息化进程的深入推进，各类学习管理系统、在线教育平台和学习分析工具广泛应用于教育教学过程，产生了海量的学习行为数据。这些数据蕴含着丰富的潜在知识，对于理解学习规律、优化教学设计和实现个性化学习具有重要价值。据教育部统计，2020 年全国高校教师开设了超过 1000 万门在线课程，每天产生的学习行为数据量呈几何级增长，但这些数据分散在不同系统中，形成了数据孤岛，难以有效整合利用，这严重阻碍了教育大数据的应用价值发挥。

为有效打破数据壁垒，显著提升数据共享程度和利用率，文献 [1] 构建了研究性学习行为记录库，实现了多源数据融合架构；文献 [2] 设计了一种基于多源数据融合的共享数据模型建模方法，结合 xAPI 数据规范，构建了一个可重用、可共享的数据模型；文献 [3] 基于大数据的高效组织与处理技术，通过异构数据自动收集、融合技术和校验技术，建立了异构多源的动态学习模型，并通过分析学科知识图谱中的关联关系，设计实现了基于知识图谱的习题推荐算法。

虽然当前已存在不少针对多源异构学习行为数据融合的研究工作，但是，由于近年来各类学习系统的快速发展，该类数据呈现出明显的多源异构性和大规模性，导致现有系统在数据处理效率上面临较大挑战。

针对上述挑战，本文设计了具有元数据层、行为数据层和上下文数据层的三层结构学习行为标准数据集，为异构数据的语义整合提供了统一框架，以解决学习行为数据多源异构性问题；在此基础上，利用 Spark 并行架构 [4] 构建了一个多源异构学习行为数据融合模型 Spark-DataFusion，该模型可利用 Spark 的 RDD 弹性分布式数据集 [5]，实现学习行为数据融合的并行计算。

1　多源异构学习行为标准数据设计

为了应对多源异构数据的复杂性，本文提出三层架构学习行为标准数据集设计方案，分层结构设计的优势是便于扩展和维护。数据标准集分为元数据层

收稿日期：2025-06-25
基金项目：河南省科技攻关项目（232102321065, 252102211064,222102240046）
第一作者简介：王崇科（1980—　　），男，河南辉县人，副教授，硕士，主要从事数字课程推荐算法研究。
通信作者简介：任刚（1978—　　），男，河南洛阳人，副教授，博士，主要从事并行算法与软件研究。

(meta-data layer)、行为数据层 (behavior-data layer) 和上下文数据层 (context-data layer)3 部分。

元数据层用于描述数据的来源、采集时间、采集工具等信息，可以为后续的数据分析提供上下文信息。*meta-data* 为元数据层数据，形式化如下：

$$meta\text{-}data = \{user\_id, source\_system, data\_type, timestamp, session\_id\} \quad (1)$$

行为数据层是标准集的核心部分，用于存储用户的学习行为数据，统一不同学习系统中行为的语义和格式。*behavior-data* 为行为数据层数据，形式化如下：

$$behavior\text{-}data = \{action\_type, action\_object, duration, score, status\} \quad (2)$$

上下文数据层于描述用户行为发生时环境信息，例如设备类型、地理位置、网络条件等。这些信息可以为行为分析提供额外维度。*context-data* 为上下文数据集，形式化如下：

$$context\text{-}data = \{device\_type, os\_type, browser\text{-}type, location, network\text{-}type\} \quad (3)$$

基于上述定义，不同学习系统的异构学习行为数据可形式化为：

$$r = (meta\text{-}data, behavior\text{-}data, context\text{-}data) \quad (4)$$

学习行为标准数据集 R 可表示为：

$$R = \{r_0, r_1, \cdots, r_{n-1}\} \quad (5)$$

其中，$r_i = (meta\text{-}data_i, behavior\text{-}data_i, context\text{-}data_i)$。

学习行为 3 层标准化数据集能够有效解决多源异构数据的整合问题，为学习行为数据融合计算奠定基础。

2　基于 Spark 的多源异构学习行为数据融合计算模型

2.1　HDFS

Hadoop 分布式文件系统 (Hadoop Distributed File System，HDFS) 是以 PC 机群为硬件基础的存储系统 [6-8]。HDFS 包含名称节点 (NameNode) 和数据节点 (DataNode) 两个角色。名称节点管理文件存储位置信息，而数据节点负责存储具体的文件数据，总体框架如图 1 所示。当客户端需要写入文件时，一个输入文件首先在客户端即被分为若干个块 (Block)，一个块通常为 128 MB 至 512 MB 之间。然后，由客户端向名称节点请求写入位置，收到写入位置后，按照名称节点指定的位置直接将数据块写入数据节点。该过程中，客户端直接与数据节点交互，该方式能够保证在并发情况下保持较高的写入效率。另外，通常情况下，一个文件块还存在 2 个副本作为备份；副本由数据节点负责写入，客户端和名称节点不直接参与，可保证系统效率。当客户端读取文件时，则先向名称节点请求文件位置信息，收到文件位置信息后，按照指定位置信息从数据节点读取文件。此过程中，名称节点仅负责位置信息的查询工作，并不直接参与数据文件读取工作，此方式能够保证在并发任务较多的时候，系统的性能不受太大影响。HDFS 集群节点之间通过 Java 套接字组播传输模式通信。该方式下，数据被组合成分组形式传播，可以有效提高传输效率。

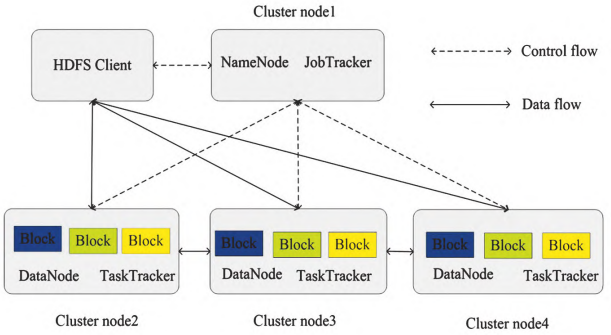

图 1　HDFS 总体框架

2.2　Spark 大数据并行计算架构

Spark 是一种基于 HDFS 的大数据并行计算模型 [9,10]，该模型是以弹性分布式数据集 (Resilient Distributed Dataset, RDD) 为核心的并行计算模型。RDD 是基于内存的只读数据集合，通过转换 (Transformation) 和行动 (Action) 两类操作完成业务逻辑。转换操作包括 map、groupByKey 和 reduceByKey 等。行动操作包括 count、reduce(func) 和 foreach(func) 等。在整个数据处理过程中，RDD 采用惰性机制，也就是说，RDD 只有碰到数据行动操作时，才会进行计算，其计算架构如图 2 所示。该模型从 HDFS 读入数据，形成 RDD，通过转换操作，结果重新写回 HDFS。在利用 RDD 进行遗传算法迭代操作时，通常采用键值对 RDD 这种数据集合形式，该数据集合每个元素都是 (*key, value*) 键值对类型，主要通过 map、reduceByKey 和 groupByKey 操作完成业务逻辑。

2.3　Spark-DataFusion 模型实现

本文利用 Spark 大数据并行计算架构实现多源异构学习行为数据融合计算模型。该模型的实现过程包括数据标准化和数据融合两个核心阶段。数据标准化阶段由 map1 完成，数据融合阶段由 map2 和 groupByKey 完成。学习系统原始数据最初存储在各

自的数据库中。这些数据库包含不同格式和结构的学习行为记录，需要通过统一的流程进行标准化和融合处理。数据标准化阶段负责从各个学习系统的数据库中读取原始数据，并将其转换为标准化的三层数据结构。数据融合阶段将标准化数据结构按照用户维度进行融合。模型并行架构如图 3 所示。

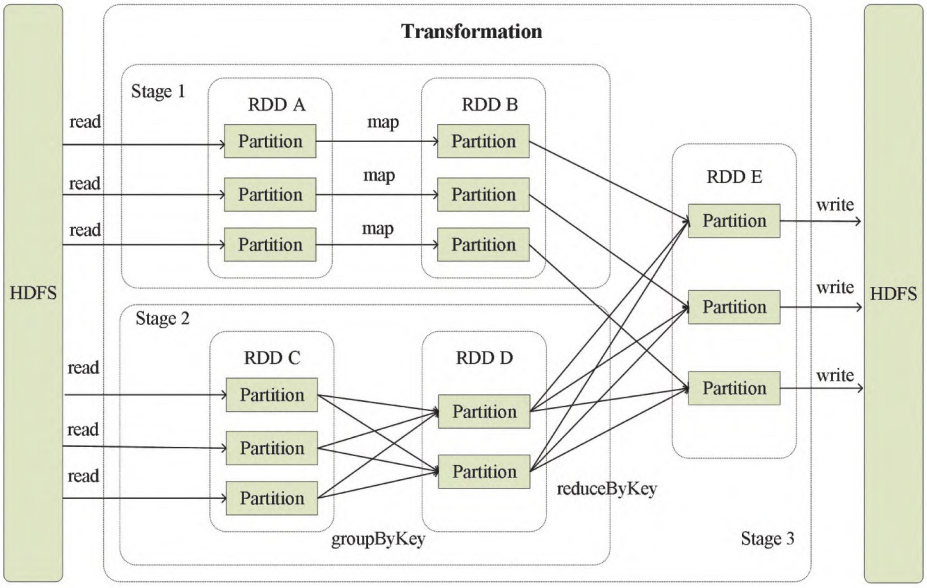

图 2 Spark 并行架构

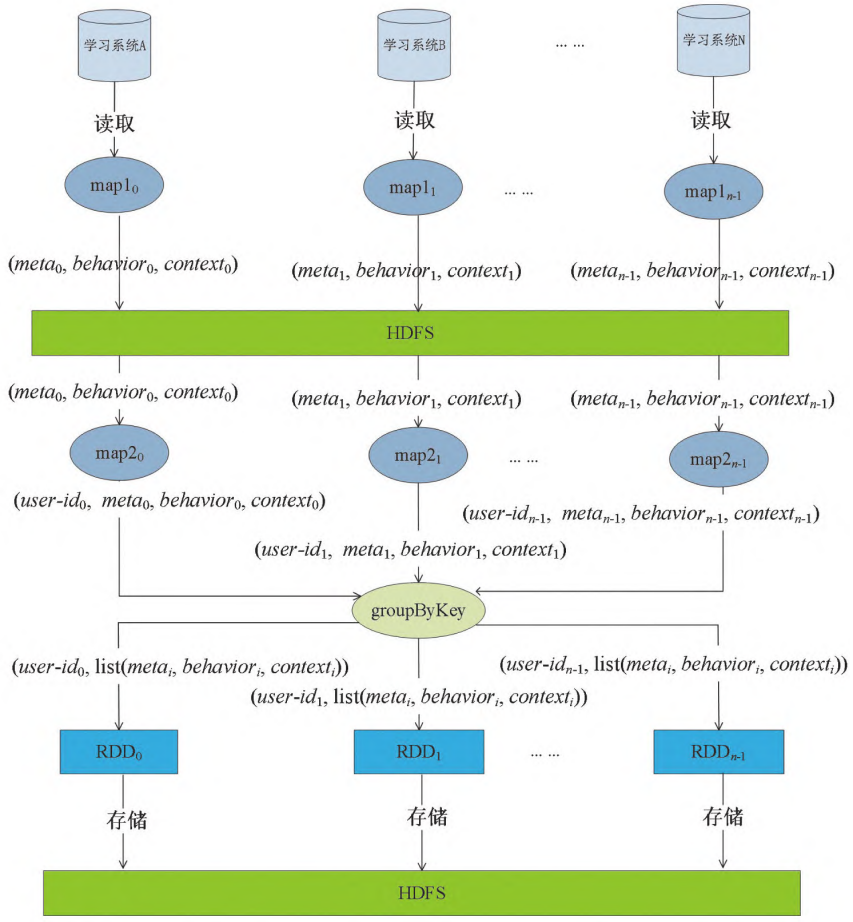

图 3 Spark–DataFusion 模型并行架构

（1）map1 实现。假定 $D_i = \{d_1, d_2, \cdots\}$ 表示学习系统 i 原始数据集，map1 首先读取原始数据集 d_i，将其转换为单值 RDD，命名为 $value_1$，形式化表示为：

$$value_1 = d_j \quad (6)$$

map1 从 $value_1$ 中分别读取 *meta*、*behavior* 和 *context* 数据集，形成标准化数据集 $value_2$，其形式化表示如下：

$$value_2 = (meta, behavior, context) \quad (7)$$

主要实现代码如下：

算法 1：map1

```
Input: value1
Output: value2
01  meta = value1.getMeta( );
02  behavior = value1.getBehaviror( );
03  context = value1.getContext( );
04  value2 = (meta, behavior, context);
05  value2.savetoHDFS( );
```

（2）map2 实现。map2 负责为数据融合准备基础数据，首先，从 HDFS 读取 $value_2$，并提取 *user-id*，生成键值对 RDD，由 $(key_3, value_3)$ 构成，其形式化表示如下：

$$key_3 = user\text{-}id \quad (8)$$

$$value_3 = (meta, behavior, context) \quad (9)$$

主要实现代码如下：

算法 2：map2

```
Input: value2
Output: (key3, value3)
01  user-id = value2.getUser–id( );
02  key3 = user-id;
03  value3 = value2;
04  Emit(key3, value3);
```

（3）groupByKey 实现。groupByKey 负责根据 *use-id* 进行数据融合，将相同用户的所有记录整合到一个 list 中，生成 (*use-id, list*(*meta, behavior, context*)) 格式数据。输入为 $(key_3, value_3)$，输出为 $(key_4, value_4)$。具体表示如下所示：

$$key_4 = key_3 \quad (10)$$

$$value_4 = ((meta_0, behavior_0, context_0), (meta_1, behavior_1, context_1), \cdots) \quad (11)$$

主要实现代码如下所示：

算法 3：groupByKey

```
Input: (key3, list(value3))
Output: (key4, value4)
01  key4 = key3;
02  value4 = list(value3);
03  value4.savetoHDFS( );
```

3 算法验证

实验采用的 Spark 集群中含 8 个计算节点，每个计算节点由 12 核 2.1 GHz 处理器、16 GB 内存和 1T 硬盘构成。为了验证算法的性能，分别进行了数据规模扩展性实验和计算节点扩展性实验，试验结果分别如图 4 和图 5 所示。

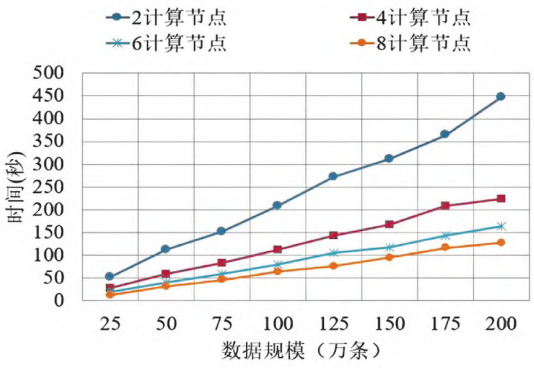

图 4 数据规模扩展性

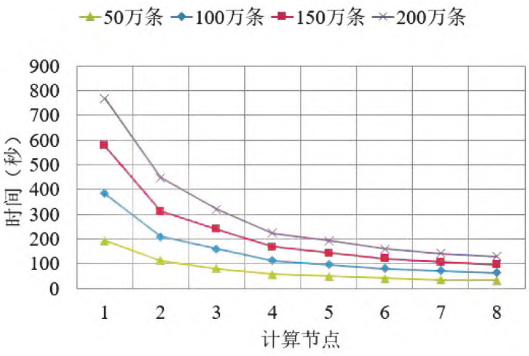

图 5 计算节点扩展性

3.1 实验 1：数据规模扩展性分析

该实验分析数据规模对计算时间的影响。数据规模从 25 万条记录开始，每次增加 25 万条，直到 200 万条，共设置 8 个数据规模节点。分别测试 2、4、6 和 8 个计算节点下的处理时间。由图 4 可以观察到：在固定计算节点数的情况下，处理时间随着数据规模的增加呈线性增长，这说明算法具有良好的数据规模扩展性；在相同数据规模下，处理时间随着计算节点数的增加而减少，计算节点数增加带来的性能提升呈现边际递减效应。

3.2 实验 2：计算节点扩展性分析

该实验分析计算时间和计算节点数的关系。实验设置了从 1 到 8 个计算节点，测试在不同数据规模（50 万、100 万、150 万、200 万）下的处理时间。由图 5 可以看出：在相同数据规模下，处理时间随着

计算节点数的增加而减少，验证了算法的并行计算效果，性能提升呈现明显的边际递减趋势；不同数据规模下的处理时间保持线性关系，说明算法在各种规模的数据处理中表现稳定。

综上所述，该算法具有良好的数据规模扩展性和计算节点扩展性，能够有效处理大规模数据，并且通过增加计算节点可以显著提升处理效率。

4 结论

本文构建了一个具有三层结构的标准化学习行为数据集，提出了基于 Spark 并行计算框架的多源异构学习行为数据融合模型。实验结果表明，该模型具有良好的数据规模扩展性和计算节点扩展性，能够有效处理大规模学习行为数据。未来可在融合算法优化、数据质量保障、实时处理能力等方面作进一步的研究。

参考文献：

［1］余明华，张治，刘小龙．基于 xAPI 的研究性学习多源数据融合研究 [J]. 现代远距离教育，2022, 30(3): 63–69.

［2］武法提，黄石华．基于多源数据融合的共享教育数据模型研究 [J]. 电化教育研究，2020, 41(5): 59–65，103.

［3］李向阳．基于异构多源学习数据的智慧教育平台研制 [D]. 杭州：浙江工商大学，2020.

［4］丁家满，李海滨，邓斌．一种基于 Spark 的频繁项集快速挖掘算法 [J]. 软件学报，2023, 34(5): 2446–2464.

［5］许利杰，方亚芬．大数据处理框架 Apache Spark 设计与实现 [M]. 北京：电子工业出版社，2020.

［6］林子雨．大数据技术原理与应用 [M]. 北京：人民邮电出版社，2024.

［7］DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107–113.

［8］CIRITOGLU H E, MURPHY J, THORPE C. HaRD: a heterogeneity–aware replica deletion for HDFS[J]. Journal of Big Data, 2019, 6(1): 94–115.

［9］任刚，李鑫，刘小杰．基于 Spark 大数据计算模型的遗传算法深度前馈神经网络训练算法 [J]. 河南工学院学报，2023,31(5): 14–22.

［10］任刚，吴长茂，魏勇．基于 Spark 大数据计算模型的多种群并行进化遗传算法 [J]. 河南工学院学报，2021, 29(3): 26–32.

Multi–source Heterogeneous Learning Behavior Data Fusion Model Based on Spark Parallel Architecture

WANG Chongke, REN Gang, WEI Yong

（School of Computer Science and Technology, Henan Institute of Technology, Xinxiang 453003, China）

Abstract: In the current process of educational informatization, learning behavior data exhibits significant multi–source heterogeneity and large–scale characteristics, posing considerable challenges to data fusion computing. This paper designs a three–layer structured learning behavior standard dataset, including a metadata layer, behavioral data layer, and contextual data layer, providing a unified framework for semantic integration of heterogeneous data. Based on this dataset, a multi–source heterogeneous learning behavior data fusion model is constructed using Spark's parallel architecture. Experimental results demonstrate that the model exhibits excellent data scalability and computational node scalability, significantly improving the processing efficiency of large–scale learning behavior data, which has positive significance for promoting deep mining of big educational data.

Key words: Spark; learning behavior data; data fusion; parallel computing

（责任编辑　王　磊）