计算机科学
COMPUTER SCIENCE

面向回收信息的线上线下多源异构数据融合系统

仇明鑫，雷帅，柳先辉，张颖瑶

引用本文

仇明鑫，雷帅，柳先辉，张颖瑶. 面向回收信息的线上线下多源异构数据融合系统[J]. 计算机科学，2024，51(11A)：240100095-7.

QIU Mingxin, LEI Shuai, LIU Xianhui, ZHANG Yingyao. Online and Offline Multi-source Heterogeneous Data Fusion System for Recycling Information [J]. Computer Science, 2024, 51(11A): 240100095-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240100095

面向回收信息的线上线下多源异构数据融合系统

仇明鑫　雷　帅　柳先辉　张颖瑶
同济大学电子与信息工程学院　上海 201804
(2233091@tongji.edu.cn)

摘　要　资源循环利用产业的废旧产品回收过程中多系统协同工作会产生大量多源异构数据,针对废旧产品线上线下回收信息难以融合并有效利用的问题,提出了一种面向回收信息的线上线下多源异构数据融合系统。首先,系统采用 Web API 接口实现线上线下多源异构数据的数据接入,通过数据解析、数据清洗及数据转换等步骤完成对多源异构数据的预处理。其次,针对现有基于聚类分析的数据融合方法在融合过程中往往还需预先指定聚类簇数的问题,提出了一种基于多目标聚类的融合方法,以在融合过程中自动确定聚类簇数。通过对预处理后的数据进行特征选择、标签编码、数据转换和归一化处理,结合多目标聚类算法完成对部分典型数据的特征提取与聚类,并对全量及增量数据进行基于欧氏距离的数据匹配。最后,系统采用了基于 MyCat 中间件及 MySQL 主从复制的分布式数据库方案,以实现融合数据的存储与共享交换。测试表明,该数据融合系统可以实现对废旧产品线上线下多源异构回收信息的数据融合及共享交换,同时,相比基于 K-Means 的数据融合方法,所提出的基于多目标聚类的数据融合方法在不同数据集上都能够自动确定最优聚类簇数,并且能够获得不差于 K-Means 融合方法的簇内紧密性和簇间分离性。

关键词:聚类;多目标优化;多源异构数据;数据融合
中图分类号　TP391

Online and Offline Multi-source Heterogeneous Data Fusion System for Recycling Information

QIU Mingxin, LEI Shuai, LIU Xianhui and ZHANG Yingyao
School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China

Abstract　In the recycling process of waste products in the resource recycling industry, a large number of multi-source heterogeneous data will be generated due to the collaborative work of multiple systems. Aiming at the problem that the online and offline recycling information of waste products is difficult to fuse and effectively use, an online and offline multi-source heterogeneous data fusion system for recycling information is proposed. Firstly, the system uses the Web API to realize the data access of online and offline multi-source heterogeneous data, and completes the pretreatment of it through the steps of data parsing, data cleaning and data conversion. Secondly, aiming at the problem that the existing data fusion methods based on clustering analysis usually need to specify the number of clusters in advance in the fusion process, a fusion method based on multi-objective clustering is proposed, which aims to automatically determine the number of clusters in the fusion process. Through feature selection, label co-ding, data conversion and normalization of the preprocessed data, combined with the multi-objective clustering algorithm, feature extraction and clustering of typical data is completed, and data matching based on Euclidean distance is performed for the total and incremental data. Finally, the system uses a distributed database scheme based on MyCat middleware and MySQL master-slave replication to realize the storage, sharing and exchange of fusion data. The test shows that the data fusion system can realize the data fusion, sharing and exchange of online and offline multi-source heterogeneous recycling information of waste products. At the same time, compared to the method based on K-Means, the proposed data fusion method based on multi-objective clustering can automatically determine the optimal cluster number on different data sets, and can obtain the compactness and separation no worse than that of the K-Means fusion method.

Keywords　Clustering, Multi-objective optimization, Multi-source heterogeneous data, Data fusion

1　引言

近年来,“大量生产、大量消费、大量废弃”的生产生活模式造成了资源的短缺和环境的恶化,严重影响到人们的生存环境[1]。而随着“十四五”规划提出“加强废旧物品回收设施规划建设,完善城市废旧物品回收分拣体系”,以及党的十九届五中全会明确要求“加快构建废旧物资循环利用体系”,废旧产品回收产业领域正不断扩张。并且随着信息化、绿色化发展的深入推进,“互联网+”也为资源循环利用产业的持续发展注入了新活力,凭借信息系统的支撑,在充分整合信息资

基金项目:国家重点研发计划(2022YFB3305802)
This work was supported by the Key Research and Development Program of China(2022YFB3305802).
通信作者:张颖瑶(zhangyingyao@tongji.edu.cn)

源的基础上高效利用信息资源，有助于突破其来源分散、回收价值低、规模小、融资难等瓶颈制约，积极推动基于大数据时代的资源循环利用产业的发展[1]。而在废旧产品回收过程中，由于多个系统的协同工作会产生大量不同来源的数据，这些结构化、半结构化的不同来源的数据组成了线上线下多源异构数据。由于这些数据的多源异构性，使得回收信息难以融合并得到有效利用。随着多源异构数据在废旧产品回收过程中的深入应用，面向废旧产品回收信息的数据融合正成为需推进的研究课题。

随着大数据分析的蓬勃发展与广泛应用，数据融合技术的研究逐渐受到重视。Xia 等[2]将联合卡尔曼滤波算法应用到配电网多源异构数据融合场景中，通过构建数据纠偏机制、最小二乘法以及拉格朗日插值法对数据进行相应的预处理，再利用联合卡尔曼滤波法将相同数据融合到一个类中以实现多源异构数据融合。针对车辆卫星导航系统无法覆盖隧道、地下车库等场景的问题，Li 等[3]采用特征匹配法建立基于众包数据的特征信息库，同时对车辆多种传感器提供的观测量以特征方式进行联合匹配。针对复杂电力系统中规模庞大、体系复杂的多源异构数据，Lin 等[4]搭建了以 Kafka 为数据输入源、Storm 为数据处理平台的实时大数据分析架构。针对智慧社区中多源异构数据难以存储、融合和共享等问题，Ku 等[5]基于 XML 格式实现多源异构数据的封装解析与传输存储，并基于 Hadoop 架构构建了大数据管理平台。针对医疗信息网络中多源异构数据的有效信息提取问题，Li 等[6]提出了一种改进的 RNN 多源融合算法来实现对医疗数据特征的深度挖掘。针对高速公路海量机电状态数据价值难以挖掘的问题，Tan 等[7]提出采用神经网络来实现对车流量、车辆速度等状态数据的特征提取及分类。针对心脑血管疾病的半结构化数据，Alhgaish 等[8]设计了一个基于数据湖技术的数据管理框架，并且采用 K-Means 聚类方法实现对具有大数据特征的分类和数值数据的融合。针对多源异构影像数据的融合问题，Hui[9]构建了一种泛化性强的深度学习模型，将深度特征学习技术运用到多源异构影像数据的提取、融合和挖掘中。

目前关于多源异构数据融合技术在废旧物资循环利用领域的应用还比较匮乏，同时现有的研究还存在着以下问题：1)不同数据源的废旧产品回收信息可能使用不同的数据格式和字段，同时还可能存在重复或冗余信息，但以联合卡尔曼滤波法和特征匹配法为代表的数据级融合方法仅将数据进行整合，在融合过程中容易忽略数据的一致性和冗余性等问题。2)废旧产品回收信息中可能存在着有待挖掘利用的关联关系，但基于大数据平台的决策级融合往往只关注决策结果，而缺乏对原始数据内在关联的挖掘，难以发现废旧产品回收信息中隐藏的关联关系。3)基于深度学习和聚类分析等方法的特征级融合可以很好地帮助挖掘数据间隐藏的关联关系，以便为后续数据分析应用提供支持。然而目前基于聚类分析的特征级融合方法在其融合过程中往往还需要预先指定聚类簇数，但未知数据集的最优聚类簇数通常难以获得。

针对以上问题，提出了一种基于多目标聚类的特征级融合方法以帮助挖掘回收信息中隐藏的关联关系，该方法能够在融合过程中自动确定最优聚类簇数，在此基础上提出了一种面向回收信息的线上线下多源异构数据融合系统，以解决废旧产品回收信息难以融合的问题。首先，系统采用 Web API 接口实现线上线下多源异构数据的数据接入，以统一管理并处理不同渠道和系统的数据。这个过程中包括将不同格式存储的数据文件进行数据解析并统一转换为 DataFrame 格式，同时对其进行数据清洗以提高数据的准确性，以及对清洗后的数据进行数据转换以确保数据的一致性；其次，该系统通过人工的方式筛选部分典型数据，并对预处理后的典型数据进行特征选择，保留有价值的数据特征以降低数据的复杂性及计算成本；随后，对选择出来的特征列进行标签编码、数据转换及归一化处理以将其转换为聚类模型的输入数据，结合多目标聚类算法完成对输入数据的特征提取与聚类，采用基于欧氏距离的数据匹配方式完成对全量数据及增量数据的类别划分；最后，为实现融合数据的存储与共享交换，提升数据库系统性能，采用了基于 MyCat 中间件及 MySQL 主从复制的分布式数据库方案。该系统旨在有效解决废旧产品回收信息的融合及有效利用问题，发挥多源异构回收数据的互补性，弥补各个回收系统中单个数据源所包含信息不完整的缺点，同时帮助发现数据中存在的隐藏关联关系，为数据分析应用提供支持。

2 系统概述

本文提出了一种面向废旧产品回收信息的线上线下多源异构数据融合系统，系统架构如图 1 所示。系统设计遵循层次化、模块化的思想，并且具有高可拓展性，在层次上可分为数据接入层、数据融合层和数据存储及共享层 3 个层次。

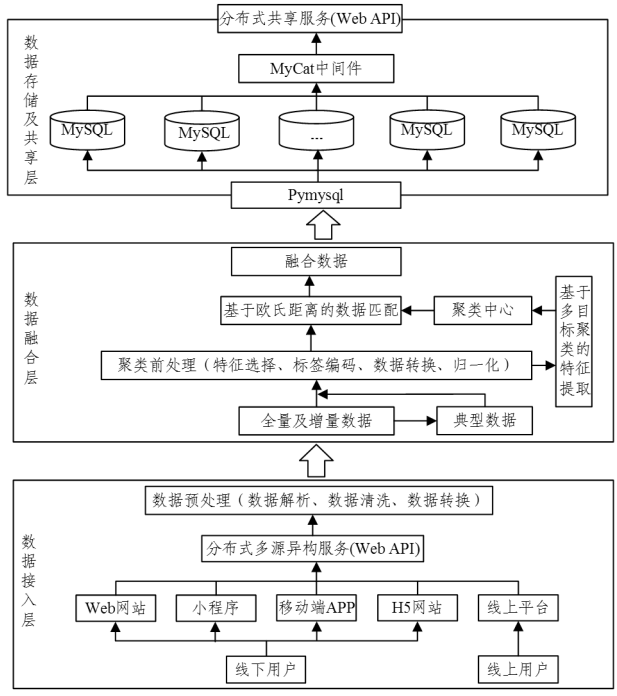

图 1 多源异构数据融合系统架构

Fig. 1 Architecture of multi-source heterogeneous data fusion system

2.1 数据接入层

数据接入层主要借助 Web API 数据接口将来自不同系统的异构数据进行整合，并解析为统一的数据格式。同时由于多个回收系统传入的回收数据通常存在重复、缺失、错误、字段不统一的问题，将数据接入到服务器后需对数据进行数据清洗及数据转换，以保证数据的准确性和一致性。

2.2 数据融合层

数据融合层包括两部分内容,即采用多目标聚类算法对业务中筛选出来的典型数据进行特征提取并聚类,以及根据聚类结果采用基于欧氏距离的数据匹配方法对该业务中的全量及增量数据进行数据匹配,并分配簇标签,最终完成数据的融合。融合旨在发挥多源异构回收数据的互补性,弥补各个回收系统中单个数据源所包含信息不完整的缺点,同时将具有相同簇标签的数据聚合在一起,以便帮助挖掘数据间隐藏的关联关系,为数据分析应用提供支持。

2.3 数据存储及共享层

数据存储及共享层的任务是采用 Pymysql 工具将融合后的数据统一以结构化数据的形式存入数据库中,并提供分布式共享服务(Web API)以提高融合数据的利用率,将数据共享给 Web 网站、移动 App、微信小程序等用户前端与运维平台,让在不同地方使用不同计算机、不同软件的用户能够读取数据并进行各种操作、运算和分析。

3 多源异构数据融合关键技术

3.1 线上线下多源异构数据的接入及预处理

由图 1 所示的系统架构可知,数据接入层先后主要可分为两个步骤,即线上线下多源异构数据的接入,以及对接入的异构数据进行数据预处理。

3.1.1 线上线下多源异构数据的接入

多源异构数据指:1)数据来源于不同的子系统;2)数据结构及传输方式不同,具有形式复杂、形态多变的特点。废旧产品回收过程中由于多系统协同工作会产生大量来源不同的数据,这些数据包含线下数据和线上数据。线下数据主要来自于线下用户使用的 Web 网站;微信小程序回收端口、回收应用移动终端 APP、H5 网站等;线上数据主要来自于企业所使用的线上产品回收平台中的回收信息,总体包括业务型数据库数据、JSON 数据、XML 数据、CSV 数据、TXT 数据和 XLSX 数据等,这些结构化、半结构化的不同来源和不同规范的数据组成了多源异构数据。

针对废旧产品回收信息的数据来源及特点,本系统采用调用 Web API 数据接口的方式来实现线上线下多源异构数据的接入。废旧产品回收信息多为半结构化数据,对于各业务数据库中已存在的结构化数据,则可以统一转换为半结构化数据后再传入 Web API。

3.1.2 数据预处理

多源异构数据的预处理是实现废旧产品回收信息数据融合的基础,为了充分发挥数据价值,提高数据质量及利用效率,首先需对接入的原始数据进行预处理。本系统的数据预处理主要分为 3 个步骤:数据解析、数据清洗与数据转换。

由于不同格式的数据文件可能包含不同的字段和结构,因此,在接入多源异构数据后首先需对数据文件进行数据解析。数据解析的任务是将传入的不同格式的数据文件(如 JSON,XML,CSV,TXT,XLSX 等)统一标准化为一致的格式(如 DataFrame),以便后续进行处理、分析以及上传至数据库等。

数据清洗有助于提高数据的质量、准确性和一致性,为后续的分析和建模提供可靠的数据基础,其主要目的是消除数据中的错误、重复、不一致和缺失的部分。通过检测重复数据并对其进行去重操作、检测并剔除与其他数据不一致的异常值或离群点来完成对错误、重复和不一致数据的消除。同时,对于部分存在缺失的数据,为了不影响数据融合的复杂度和准确性,将直接忽视不作为融合数据的原始数据,或交予客服处理。

数据转换是将不同数据源的字段标准化为具有一致格式和单位的过程。这可能涉及将数据从一种数据类型、度量单位或日期格式转换为另一种,以确保数据在整合后的一致性。

3.2 基于多目标聚类的数据融合

简单地将多源异构数据整合起来并对其进行预处理,是特征级数据融合的第一步。系统采用了一种基于多目标聚类的数据融合方法,即在数据融合层中,采用多目标聚类算法对预处理后的部分典型数据进行特征提取并聚类,以更好、更准确地反映目标事物属性。随后,采用基于欧氏距离的数据匹配方式完成对全量数据及增量数据的类别划分,最终实现多源异构数据的融合。

3.2.1 基于多目标聚类的特征提取

考虑到全量数据非常庞大,对全量数据进行聚类可能需要大量且不必要的计算资源和时间,并且在全量数据的基础上还有增量数据。因此,系统采用对部分典型数据进行聚类的方案,即在了解数据语义的前提下,从全量数据中筛选出部分典型、具有代表性的数据进行聚类,以减少计算开销。

在进行多目标聚类前,需要对数据进行聚类前的预处理,以便将数据转换为聚类模型的输入,这个过程包括特征选择、标签编码和数据转换以及归一化处理。首先,针对具体业务场景挑选出部分典型数据,基于回收业务和数据的语义理解对其进行特征选择,筛选其中用于聚类的特征列以降低数据的复杂性及计算成本。其次,对筛选出的特征列进行标签编码,将字符串数据按特定规则编码为数字,同时对日期格式等类型的特征数据进行数据转换。最后,对数据进行归一化处理,以消除特征列之间的量纲影响,避免奇异样本数据导致的不良影响。

为将相似的数据融合到一个类中,系统采用多目标聚类算法来对数据集进行特征提取并获得聚类中心。与仅采用单个聚类标准来对对象进行划分的单目标聚类算法不同,多目标聚类算法可通过同时优化多个聚类评价标准来获得聚类结果,这样可以增加算法对大部分聚类结果的鲁棒性[10-12]。

1) 目标函数

现有多目标聚类算法通常采用可以度量簇内紧密性和簇间分离性的指标作为目标函数,但往往需要预先指定聚类簇数 K[13-16]。在对未知的数据集进行聚类时,确定聚类簇数 K 往往非常麻烦,尤其当数据集较大或聚类数很多时,重复运行聚类算法以得到最佳聚类簇数 K 的方法十分低效。已有研究提出了一种自动确定聚类簇数 K 的多目标聚类方法[17-18],该方法中 K 作为决策变量,与另一种有相互冲突的聚类评价指标共同作为多目标优化算法的目标函数,最终可以获得一个最优聚类簇数 K 及其所对应的聚类中心解。

簇内距离平方和 Sum of Squared Distance(SSD)指簇内数据点与其簇中心之间距离的平方和,其值越小,代表着每个簇内样本越相似。因此,一般 K-Means 算法的思想就是求解能够让 SSD 最小化的簇中心。同时,K 的减小代表着簇的数量减少。每个簇中的数据点增多,这使得每个簇内的数据点

更加紧密，簇内样本更加相似。然而，两者往往不能同时达到最小，SSD 通常随着 K 的增加而减小，在极端情况下，即 K 与数据点的数量相同时，SSD 减小到零。因此，K 代表簇间分离性，SSD 代表簇内紧密性。为了自动确定 K 并获得最优聚类效果，选取聚类簇数 K 和簇内距离平方和 SSD 作为目标函数。但为了保证每个 K 下都有一个解决方案，采用 SSD 的变体[19]作为其中一个目标函数 f_1，K 和 SSD 值越小表示聚类效果越好，如式(1)所示：

$$Min\ F(x)=\{f_1(x)=(1-\exp^{-1\cdot SSD})-K, f_2(x)=K\} \tag{1}$$

SSD 的计算方式如下：

$$SSD=\sum_{r=1}^{K}\sum_{x_i\in C_r}\| x_i-u_r \|^2 \tag{2}$$

$$u_r=(u_r^1,u_r^2,\cdots,u_r^d)$$

其中，C_r 表示第 r 簇中数据点的集合，x_i 表示数据点，$u_r=(u_r^1,u_r^2,\cdots,u_r^d)$ 表示第 r 簇的聚类中心，d 表示数据点的维度。

2）编码方式及初始种群设置

进化算法一般将种群表示为一个 $N\cdot n$ 的矩阵，代表种群中有 N 条染色体，并且每条染色体的长度为 n。采用实数编码的方式将聚类中心编码进染色体，为了处理不同 K 的情况，采用统一长度的染色体，并使 $n=d\cdot K_{\max}+1$，其中 $K_{\max}$ 为最大聚类数，并且 K 的取值范围默认为 $[2,K_{\max}]$，同时将 K 编码进染色体的最后一个基因位。算法首先随机初始化一个包含 N 条染色体的种群，代表了对问题的 N 种解，初始化后，每条染色体都被分配了一个随机的 K，因此，在搜索过程中，只有前 $d\cdot K$ 个基因位上的值被作为染色体的决策变量。此外，为了避免无效空簇，即某一簇中没有数据点附着的情况，随机从数据集中选择 K 个不同的数据点作为该染色体的前 $d\cdot K$ 个基因，剩余 $d\cdot(K_{\max}-K)$ 个基因则由输入数据集的上界和下届限定。如图 2 所示，当 $K_{\max}=4$，$d=3$，并且该染色体被随机分配为 $K=3$ 时，只有前 9 个基因位上的值是该染色体对聚类中心真实有效的编码，并且由数据集中 3 个不同数据点在空间中的坐标组成。

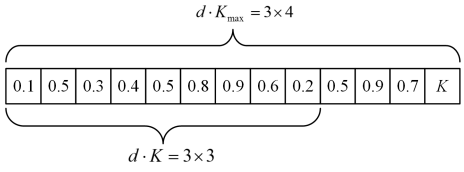

图 2 染色体编码示意图

Fig. 2 Schematic diagram of chromosome coding

3）多目标聚类算法流程

本文采用的多目标聚类算法流程如算法 1 所示，算法采用具有较好鲁棒性和简便性的多目标优化框架 NSGA-II[20]作为优化器。

算法 1 多目标聚类算法

输入：数据集 Data，最大聚类数 K_{max}，种群规模 N，最大迭代次数 maxGen，交叉概率 C_r，变异概率 P_m，交叉分布指数 η_c，变异分布指数 η_m

输出：Pareto 解

1. 随机初始化具有 N 条染色体的初始种群 P，每一条染色体代表一个聚类中心解决方案，每个染色体被随机分配一个范围内 K；
2. for Gen←1 to maxGen do
3. 通过模拟二进制交叉和多项式变异[21]的方式从种群 P 中生成数量为 N 的子代种群 Q；
4. 种群 P 与子代种群 Q 合并为数量为 2N 的种群 PQ；
5. 计算种群 PQ 中每条染色体的 f_1 值，并根据 K 和 f_1 值计算其拥挤距离和非支配排序等级；
6. 依据精英保留策略[21]从 PQ 中选取较好的 N 条染色体组成下一代种群 P；
7. end
8. 从种群 P 中选择一组非支配解作为 Pareto 解。

算法 1 中，对于通过交叉操作所产生的子代个体，其最后一个基因位上的 K 由其父代个体遗传，而变异操作仅对前 $n-1$ 个基因位上的基因进行变异。迭代结束后，算法将从最后一代种群 P 中选择出一组非支配解作为 Pareto 解。

4）最终解选取策略

多目标聚类算法终止后，需要从 Pareto 解中确定一个最终优化解。通常情况下采用“肘部法”来确定最优聚类结果，即在 SSD 与 K 的关系图中，选择斜率变化最大的“拐点”作为最佳 K[22]。但由于某些数据集中可能不止一个这样的拐点，或者根本没有拐点，从而增加了确定最优解的难度。为了帮助挖掘回收信息中相似数据间隐藏的关联关系，在聚类的过程中同一簇内的数据应更加紧密，不同簇间的数据差异性更大，因此，采用能够同时评估簇内紧密性和簇间分离性的 DB (Davies-Bouldin)指数来完成最终解的选取。

$$DB=\frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}R_{ij} \tag{3}$$

其中，$R_{ij}=\frac{s_i+s_j}{d_{ij}}$；$s_i$ 为第 i 簇内样本到其簇中心的平均欧氏距离，其值越小代表簇内越紧密；d_{ij} 为第 i 簇和第 j 簇的簇中心欧氏距离，其值越大代表簇间越分离；数据点 x 与 y 之间欧氏距离 $D(x,y)$ 的计算式如下：

$$D(x,y)=\sqrt{\sum_{i=1}^{d}(x_i-y_i)^2} \tag{4}$$

其中，d 表示数据点 x 的维度。可见，DB 越小代表算法所生成的聚类结果具有更好的簇内紧密性和簇间分离性，从而有助于更好地挖掘相似数据隐藏的关联关系。因此，采用选取 DB 最小值所对应解的方式能够从 Pareto 解中选取聚类效果最好的解作为最终优化解。

3.2.2 基于欧氏距离的数据匹配

在完成对部分典型数据的多目标聚类并得到最优聚类中心的结果后，需经过多次实验，并在理解业务和数据语义的基础上挑选出最优的聚类中心结果。由于算法 1 为基于欧氏距离的聚类方法，因此，为了节省计算开销，对于全量数据及增量数据的数据融合，可通过对其做同样的聚类前处理，并计算每个数据点在空间中与各个聚类中心的欧氏距离，将其分配到距离最近的簇中并划分簇标签，以此完成全量数据及增量数据的数据匹配。

3.3 数据存储与共享交换

实现融合数据的存储与共享交换，是对多源异构回收信息进行数据融合的最终目的。通过对废旧产品线上线下多源异构数据进行数据预处理及数据融合，并构建废旧产品回收信息数据框架，最终形成统一的数据存储格式或存储方式。而数据库的选择对融合后回收信息的管理十分重要，目前常见的数据库有 Oracle，HBase，SQL Server，Redis，PostgreSQL，MySQL 等，其对比分析如表 1 所列。通过对比发现，每种数据库技术均有其优势和适用场景。

其中 MySQL 是一种流行的关系型数据库管理系统,具有扩展方便、性能可靠、操作简单等特点。它使用了索引功能,提高了系统性能,同时可在 Windows 与 Linux 等多种操作系统上运行,种种优势使得 MySQL 成为当下最流行的数据库管理系统之一。因此,为满足融合后回收信息的存储效率及查询效率等要求,采用适用于结构化数据场景的 MySQL 数据库技术,并将融合后的数据统一以结构化数据的形式存储起来。

表 1 常见数据库对比分析

Table 1 Comparative analysis of common databases

数据库名称	Oracle	HBase	SQL Server	Redis	PostgreSQL	MySQL
数据库类型	关系型	列式型	关系型	单键型	关系型	关系型
结构化数据	支持	部分支持	支持	不支持	支持	支持
是否支持事务	支持	部分支持	支持	不支持	支持	支持
可拓展性	高	高	中等	高	高	高
是否开源	否	是	否	是	是	是
适用场景	对数据安全较为重视的场景	具有数据设计统计场景	专有软件及海量数据管理场景	数据缓存及低延时场景	专有软件及海量数据管理场景	结构化数据场景

为构建高可用、可扩展、负载均衡的 MySQL 数据库系统,采用基于 MyCat 中间件及 MySQL 主从复制的分布式数据库方案,以提高融合数据存储与共享交换的效率和可靠性。如图 3 所示,在该数据库方案中,Web API 使用 Pymysql 工具连接 MyCat 节点,并借助 MyCat 中间件来实现对 MySQL 多个“一主一备二从”架构的读写。这种可扩展的“主备从”架构通过二进制日志的方式实现架构中多个 MySQL 数据库之间的数据同步。在第一个“主备从”架构中,主数据库节点(M_1)负责融合数据的写入,从数据库节点(S_1,S_2)负责融合数据的读取,备用数据库节点(M_2)通过在闲时负责融合数据的读取,并且在主数据库节点宕机时替代主数据库完成融合数据的写入,以此确保 MySQL 数据库的高可用性。在 MyCat 中间件的管理下,多个主数据库之间可实现分库分表以提升数据库性能,同时可根据负载情况动态调整访问的主数据库节点以实现负载均衡。此外,多个从数据库与闲时的备用数据库之间可实现读操作的负载均衡。

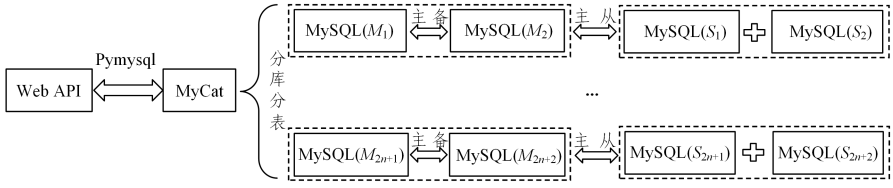

图 3 基于 MyCat 中间件及 MySQL 主从复制的数据库方案

Fig. 3 Database solution based on MyCat middleware and MySQL master-slave replication

4 实验测试

本章将检验系统实际应用效果,并依次介绍系统环境部署、实验数据集、数据融合效果及系统核心功能测试。

4.1 系统环境部署

面向回收信息的线上线下多源异构数据融合系统的环境部署为 2 个“主备从”节点,即 2 个主节点、2 个备用节点、4 个从节点。8 个节点通过在一台物理机上使用虚拟化技术进行部署,并且每个节点所采用的硬件环境均为 Intel Core i7-9700 处理器,2 GB 内存,所采用的软件环境均为 CentOS7 64 位和 MySQL8.0。其中一个主节点额外采用 Python3.9,MyCat1.6.7.3,Scikit-Learn1.2.2 以运行融合系统中的 Web API。

4.2 数据集准备

实验采用自构建的家用电器回收订单数据集,数据来源于“拍拍二手交易”小程序,包括冰箱、空调、彩电、冷柜、热水器、洗衣机 6 种,以二手交易数据来模拟构建废旧产品回收订单数据集。将以上数据集作为干净数据集,并主动地添加少量重复、异常及缺失数据构成输入数据集。同时为各类输入数据集分别构建 JSON,XML,CSV,TXT,XLSX 5 种格式的数据文件。采用随机的方式从各类别家电回收订单数据集中抽取 40%的干净数据样本作为典型数据集,各数据集具体信息如表 2 所列。

表 2 家电回收订单数据集及典型数据集

Table 2 Household appliance recycling order dataset and typical dataset

电器类别	输入数据集样本总数	干净数据集样本数	重复、异常及缺失数据样本数	典型数据集样本数	特征维度
冰箱	420	400	20	160	6
空调	531	511	20	204	6
彩电	912	892	20	357	7
冷柜	545	525	20	210	6
热水器	822	802	20	321	6
洗衣机	539	519	20	208	6

4.3 数据融合效果表现

目前已有研究多采用 K-Means 等单目标聚类方法来实现相似数据的类别划分[8,23]。为验证多目标聚类方法的数据融合效果,采用内部评估指标 DB 指数(见式(3))来对两者的聚类效果进行评估,其值越小代表聚类效果越优,即具有更好的簇内紧密性和簇间分离性,更有助于挖掘回收信息中隐藏的关联关系。

针对基于多目标聚类的数据融合方法,设置 $K_{\max}=15$,种群规模 $N=100$,交叉概率 $C_r=1.0$,变异概率 $P_m=1/(n-1)$,交叉分布指数 $\eta_c=15$,变异分布指数 $\eta_m=20$,最大迭代次数 $maxGen=1\,000$。对于 K-Means 聚类方法,采用 Scikit-Learn1.2.2 中 K-Means 算法包的默认参数,通过多次执行聚类操作并选择最小 DB 指数所对应解的方式来确定最优 K。

表3列出了两者在6个典型数据集上的最优 K 及对应DB值，可以发现，在单次运行就能够自动确定聚类簇数 K 的同时，所采用的融合方法在各个典型数据集上都能获得不差于 K-Means 融合方法的簇内紧密性和簇间分离性。

表3 K-Means方法与本方法在6个典型数据集上的融合效果对比

Table 3 Comparison of fusion effects between K-Means method and our method on six datasets

指标	融合方法	冰箱	空调	彩电	冷柜	热水器	洗衣机
DB	K-Means	0.7318	0.8324	0.5337	0.7710	0.3971	0.5308
	本方法	0.7318	0.8324	0.5337	0.7575	0.3971	0.5308
K	K-Means	2	2	2	4	8	5
	本方法	2	2	2	3	8	5

4.4 系统核心功能测试

表4列出了数据融合系统核心功能的测试步骤、预期结果和测试结果，系统核心功能均能实现。通过在不同节点中上传一种不同家电类别、随机数据格式的数据文件，并在数据库中获得所有家电类别的准确一致的干净数据，验证了该系统能够实现多源异构数据融合，发挥多源异构回收数据的互补性。同时由于多目标聚类算法仅对部分典型数据进行聚类，所得到的聚类中心用于完成基于欧氏距离的数据匹配，避免了对全量及增量数据的多目标聚类过程而产生的巨大计算开销，因此两个测试步骤所涉及的数据上传及数据查询功能平均完成时间不超过50ms，基本满足数据融合需求。

表4 多源异构数据融合系统核心功能测试

Table 4 Basic function test of multi-source heterogeneous data fusion system

测试步骤	预期结果	测试结果
随机挑选6个节点并在每个节点中调用该Web API上传一种家电类别的数据文件，其数据格式为JSON，XML，CSV，TXT，XLSX 5种格式中随机的一种	成功在不同节点中上传不同格式的数据文件，并在数据库中获得所有家电类别的准确一致的干净数据	各项功能均符合预期结果
在不同节点调用该Web API并按索引查询融合后的数据	成功查询到不同节点不同数据库中的信息，查询数据信息无误，无重复、异常及缺失数据，且每一个数据样本都带有一个簇标签	

结束语 针对废旧产品回收过程中由于多个系统协同工作而产生的多源异构数据难以融合并得到有效利用的问题，本文提出了一种面向回收信息的线上线下多源异构数据融合系统。

该系统首先采用Web API数据接口实现线上线下多源异构回收信息的接入，同时对接入的数据进行数据解析、数据清洗与数据转换等预处理；随后采用了一种基于多目标聚类的数据融合方法，先对筛选出的典型数据进行聚类并得到聚类中心后，再对全量数据及增量数据进行基于欧氏距离的数据匹配；最后采用基于MyCat中间件与MySQL主从复制的数据库架构实现融合数据的存储与共享交换。通过实验测试的方式验证了该数据融合系统的有效性，同时测试结果表明该数据融合系统在融合性能上能够获得不差于 K-Means 融合方法的簇内紧密性和簇间分离性，并且避免了重复多次运行以确定最优 K 的低效过程。

该系统虽能实现数据融合功能，但是所采用的融合方法还有很多方面值得深入研究。在后续研究中，如何利用ETL、大数据计算框架等相关技术进行数据抽取和计算以提高废旧产品线上线下回收信息融合的效率将是研究的重点。此外，针对多目标聚类算法中的目标函数策略，除了采用SSD作为其中一个目标函数外，后续还可以针对不同数据集，通过实验对比的方式选用其他更适合的评估指标。

参 考 文 献

[1] DU H Z, LV Z, SONG S W, et al. The Development Trends of International Resource Recycling Industry and China's Response during the 14th Five Year Plan Period under the "Double Carbon" Goal [J]. Macroeconomic Research, 2022(7): 120-128.

[2] XIA W, CAI W T, LIU Y, et al. Multi source heterogeneous data fusion in distribution networks based on joint Kalman filtering [J]. Power System Protection and Control, 2022, 50(10): 180-187.

[3] LI W, WEI D Y, LU Y, et al. Research on Vehicle Autonomous Location Method Based on Heterogeneous Feature Information Matching [J]. Navigation Location and Timing, 2019(3): 75-81.

[4] LIN Y, CHEN R C, JIN T. Multi source heterogeneous data fusion technology for complex information systems [J]. China Testing, 2020, 46(7): 1-7, 23.

[5] KU X B, ZHANG H L, YANG S. Data Management Platform of Smart Community Based on XML Format Fusion of Multi-Source Heterogeneous Data [J]. Electric Power Survey and Design, 2023(8): 1-5, 17.

[6] LI L, WANG W. Network heterogeneous information integrated management system based on improved RNN multi-source fusion algorithm [J]. Journal of Xi'an University of Engineering, 2023, 37(6): 145-152.

[7] TAN J D, LI B, LIU C Y, et al. Research on the fusion processing method of multi-source electromechanical state data for highways [J]. Highway, 2023, 68(8): 275-281.

[8] ALHGAISH A, ALZYADAT W, AL-FAYOUMI M, et al. Preserve quality medical drug data toward meaningful data lake by cluster[J]. International Journal of Recent Technology and Engineering, 2019, 8(3): 270-277.

[9] HUI G B. A deep learning based multi-source heterogeneous data fusion method [J]. Modern Navigation, 2017, 8(3): 218-223.

[10] HANDL J, KNOWLES J. Multi-objective clustering and cluster validation[J]. Multi-objective Machine Learning, 2006, 16(21): 21-47.

[11] JOSÉ-GARCÍA A, GÓMEZ-FLORES W. Automatic clustering using nature-inspired metaheuristics: A survey[J]. Applied Soft Computing, 2016, 41: 192-213.

[12] HANDL J, KNOWLES J. Evidence accumulation in multiobjective data clustering[C]// International Conference on Evolutionary Multi-Criterion Optimization. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 543-557.

[13] BANDYOPADHYAY S,MUKHOPADHYAY A,MAULI K U. An improved algorithm for clustering gene expression data [J]. Bioinformatics,2007,23(21):2859-2865.

[14] FACELI K,DE SOUTO M C P,DE ARAUJOD S A,et al. Multi-objective clustering ensemble for gene expression data analysis[J]. Neurocomputing,2009,72(13/14/15):2763-2774.

[15] MUKHOPADHYAY A,MAULIK U,BANDYOPADHYAY S. An interactive approach to multiobjective clustering of gene expression patterns[J]. IEEE Transactions on Biomedical Engineering,2012,60(1):35-41.

[16] MAULIK U,MUKHOPADHYAY A,BANDYOPADHYAY S. Combining pareto-optimal clusters using supervised learning for identifying co-expressed genes[J]. BMC Bioinformatics,2009,10(1):1-16.

[17] GUPTA A,ONG Y S,FENG L. Multifactorial evolution:toward evolutionary multitasking[J]. IEEE Transactions on Evolutionary Computation,2015,20(3):343-357.

[18] OMIDVAR M N,LI X,MEI Y,et al. Cooperative co-evolution with differential grouping for large scale optimization[J]. IEEE Transactions on Evolutionary Computation,2013,18(3):378-393.

[19] WANG R,LAI S,WU G,et al. Multi-clustering via evolutionary multi-objective optimization[J]. Information Sciences,2018,450:128-140.

[20] DEB K,PRATAP A,AGARWAL S,et al. A fast and elitist multiobjective genetic algorithm:NSGA-II[J]. IEEE Transactions on Evolutionary Computation,2002,6(2):182-197.

[21] SRINIVAS N,DEB K. Muiltiobjective optimization using nondominated sorting in genetic algorithms[J]. Evolutionary Computation,1994,2(3):221-248.

[22] HANCER E,KARABOGA D. A comprehensive survey of traditional,merge-split and evolutionary approaches proposed for determination of cluster number[J]. Swarm and Evolutionary Computation,2017,32:49-67.

[23] MEI W J,ZHENG J,JIN J,et al. Multi sensor asynchronous information fusion method based on sliding clustering [J]. Journal of Instrumentation,2022,43(6):109-117.

QIU Mingxin, born in 2000, postgraduate. His main research interests include machine learning and big data.

ZHANG Yinyao, born in 1984, Ph.D, associate professor. Her main research interests include machine learning and big data.