

基于粗糙集的多源异构网络数据智能整合方法

王焕鑫¹ 邱帅欣¹

WANG Huanxin QIU Shuaixin

摘要

在多源异构网络数据整合领域, 高效融合分散且异构的数据对提升分析质量至关重要。由于数据来源广泛、结构差异显著及语义模糊性, 导致难以准确识别实体并建立鲁棒的整合规则, 造成整合结果完整度低、大量有效信息丢失的严重后果。为此, 文章提出了一种基于粗糙集的多源异构网络数据智能整合方法。通过定义多维度元路径表达式并计算 PathSim 相似度, 识别出不同类别的数据实体, 初步判断数据的属性特征。经过标准化处理后, 采用粗糙集挖掘数据整合规则, 建立不确定性的模糊边界, 进而通过规则相似度的计算, 实现多源异构网络数据的智能整合。实验结果表明, 所提方法在 8 类真实接口数据构成的复杂环境中, 整合完整度达到 98%, 丢失数据量平均降低至 36.8 条, 为解决多源异构网络数据整合中的不确定性与信息丢失问题提供了一种有效方案, 具有重要的理论意义与工程应用价值。

关键词

数据整合; 多源异构数据; 异构网络数据; 多源异构数据整合; 粗糙集

doi: 10.3969/j.issn.1672-9528.2026.05.009

0 引言

随着互联网技术的发展, 网络所产生的数据量呈现指数级增长趋势。庞大的数据量给数据分析过程造成极大困难。尤其在多源异构网络数据的分析中, 难以实现不同来源网络数据的融合与分析。如果仅对单一维度的数据进行分析, 难以捕捉数据的真实特征, 覆盖度较差。因此, 为了提高数据分析的有效性, 需要采用更可靠的手段, 对多源异构网络数据进行整合, 保障数据分析结果的质量, 避免单一数据源分析所存在的样本偏差问题。为了解决这一问题, 领域内引起了广泛的讨论, 很多研究人员提出了数据整合方法。

当前领域内针对数据整合问题已开展了大量研究, 但现有方法仍存在明显局限性。李平等^[1]提出了基于知识图谱和 LLM 的数据整合方法。该方法在单一数据源上构建数据样本初始知识图谱, 并搭建相应 LLM 模型, 通过处理数据样本分布不平衡问题, 利用 LLM 模型确定整合参数, 用于处理存在数据样本缺失情况的整合任务。然而, 受数据来源广泛影响, 单一数据源难以涵盖所有相关实体与关系, 致使初始知识图谱不完整, 影响实体准确识别。师瑶等^[2]提出了基于图 Transformer 模型的数据整合方法。该方法先处理多源异构数据的异质性并去除噪声干扰, 再将处理后的数据样本输入图 Transformer 模型进行特征捕捉以实现整合。但由于数据结

构差异显著, 不同数据源在格式、编码方式等方面各不相同, 处理异质性的过程难以完全消除差异, 导致后续特征捕捉时难以准确识别实体。刘晓宇等^[3]提出了基于概率 - 非概率样本的数据整合方法。该方法在大数据技术支持下, 推断数据原始属性特征并量化计算整合倾向得分, 结合概率 - 非概率样本抽样设置 Adaptive Lasso 参数, 在该参数下完成整合。然而, 该方法在量化计算整合倾向得分时, 因对数据语义理解不准确, 导致得分计算出现偏差, 进而影响了 Adaptive Lasso 参数的设置合理性, 最终使得整合规则不够鲁棒。宋超等^[4]提出了基于大数据与聚类算法的数据整合方法。该方法采用大数据技术集成原始数据样本集合, 利用聚类算法修正异常值, 在适应度函数指导下匹配相似数据完成整合, 并对整合后的数据有序排列, 可视化效果较好。但在利用大数据技术集成原始数据样本时, 由于数据来源广泛, 其中包含大量噪声与错误信息, 导致集成过程难以有效消除这些干扰, 从而影响后续实体的准确识别。

考虑到上述研究人员所提出的数据整合方法无法实现有效的数据整合, 存在数据缺失情况, 整合后的数据完整度不高, 为解决这一问题, 本研究提出了一种基于粗糙集的多源异构网络数据智能整合方法。

1 基于粗糙集的多源异构网络数据智能整合方法设计

1.1 多源异构网络的数据实体识别与标准化处理

多源异构网络数据具有来源广泛、结构多样的特点, 导

1. 郑州科技学院信息工程学院 河南郑州 450064

致同一实体的数据分散且格式不一致。针对这一问题,本研究通过分析数据在多源异构网络中的元路径特征和网络拓扑结构特性,从而识别具有相同实体属性的多源异构数据,并对这些数据进行标准化处理,为后续数据分析和处理奠定基础。

为有效处理多源异构网络数据,建立元路径,用公式表示为^[5]:

$$P(x^i) = \{U^i \rightarrow T^i \rightarrow D^i \rightarrow X^i\} \quad (1)$$

式中: $P(x^i)$ 表示多源异构网络数据集; x 表示第 i 条数据的元路径; U^i 表示数据所属的多源异构网络节点; T^i 表示数据时间; D^i 表示数据地理位置; X^i 表示数据内容。

在集成的多源异构网络数据集中,为判断数据是否属于同一实体,需分析不同数据之间元路径的相似度。基于 PathSim 算法,定义元路径相似度的计算方法为^[6]:

$$S(x^i, x^j) = \frac{2 \times \left| \left\{ P(x^i \sim x^j) : P(x^i \sim x^j) \right\} \right|}{\left| P(x^i \sim x^i) \right| + \left| P(x^j \sim x^j) \right|} \quad (2)$$

式中: x^j 表示数据集 x 中的第 j 条多源异构数据; $S(x^i, x^j)$ 表示数据 x^i 、 x^j 两者之间的相似度; $P(x^i \sim x^j)$ 表示 x^i 与 x^j 之间的元路径实例数量; $P(x^i \sim x^i)$ 、 $P(x^j \sim x^j)$ 分别表示多源异构数据 x^i 、 x^j 自身所连接的元路径实例数量。

通过式(2)的计算,可识别出多源异构网络中属于同一实体的数据信息,实现对多源异构网络数据属性特征的初步判断。在此基础上,为避免多源异构网络数据格式不一致问题对后续数据整合过程造成影响,采用归一化公式对其进行标准化处理。用公式表示为:

$$\hat{x}^i = \frac{x^i - \mu(x^i, x^j)}{\sigma(x^i, x^j)} \quad (3)$$

式中: $\hat{x}^i$ 表示标准化的多源异构网络数据; μ 表示属于同一实体的数据均值; σ 表示数据标准差。

经过标准化处理后,将消除了格式等差异的多源异构网络数据集集成到统一的数据空间中,并依据数据的特征初步划分为不同的数据实体类别,为后续的数据整合工作奠定基础。

1.2 基于粗糙集的多源异构数据整合规则挖掘

在完成多源异构网络数据的实体识别与标准化处理,将数据集集成到统一空间并初步分类后,鉴于此类数据来源广泛、结构多样,存在大量冗余属性,这些冗余属性既增加了数据处理复杂度,又可能干扰对数据本质特征的把握,而粗糙集理论能有效处理不确定性和不精确性信息,可深入分析属性间关系以挖掘智能整合规则,故本研究建立粗糙集数据整合规则挖掘的决策信息系统,用公式表示为^[7]:

$$\begin{cases} R = (Y, C, F, V, G) \\ Y = \{\hat{x}^1, \hat{x}^2, \dots, \hat{x}^i, \dots, \hat{x}^n\} \\ C = \{a^1, a^2, \dots, a^i, \dots, a^n\} \\ G = Y \times C \rightarrow V \\ a^i \in C \cup F \end{cases} \quad (4)$$

式中: R 表示数据整合规则挖掘的粗糙集决策信息系统; Y 表示标准化的多源异构网络数据样本集合; C 表示数据样本的条件属性; a 表示条件属性的值域; F 表示信息函数, V 表示属性值域集合; G 表示多源异构网络数据样本的决策属性集合; n 表示数据样本量。

在构建的决策信息系统中,分析数据样本条件属性的传递性等特征,挖掘数据中的冗余属性,进行数据样本条件属性的约简。条件属性约简的过程用公式表示为:

$$\begin{cases} a_{\uparrow}(\hat{x}^i) = \{\hat{x}^i \in R : [\hat{x}]_B \cap Y \neq \emptyset\} \\ a_{\downarrow}(\hat{x}^i) = \{\hat{x}^i \in R : [\hat{x}]_B \subseteq Y\} \end{cases} \quad (5)$$

式中: $a_{\uparrow}(\hat{x}^i)$ 和 $a_{\downarrow}(\hat{x}^i)$ 分别表示数据样本条件属性的上、下近似集合; B 表示任意非空属性子集。

在约简后的数据样本条件属性基础上,挖掘多源异构网络数据在样本空间中的属性特征,由此挖掘出数据智能整合规则的正域、负域以及边界域^[8]。分别用公式表示为:

$$E(R) = \bigcup_{Y \in Q} a_{\downarrow}(\hat{x}^i)Y \quad (6)$$

$$N(R) = G - \bigcup_{Y \in Q} a_{\downarrow}(\hat{x}^i)Y \quad (7)$$

$$A(R) = \bigcup_{Y \in Q} a_{\uparrow}(\hat{x}^i)Y - \bigcup_{Y \in Q} a_{\downarrow}(\hat{x}^i)Y \quad (8)$$

式中: $E(R)$ 表示数据整合规则的正域; Q 表示数据样本条件属性的不确定性分类; $N(R)$ 表示整合规则负域; $A(R)$ 表示整合规则边界域。

在基于 Q 的不确定性分类下,建立数据条件属性的模糊边界,形成近似整合的模糊化概念,避免精确边界整合规则导致的数据丢失问题。

1.3 规则相似度计算下的多源异构网络数据智能整合

根据粗糙集所挖掘出的多源异构网络数据整合规则,本研究面向网络中待整合的多源异构数据,通过分析数据条件属性与整合规则之间的相似度,实现数据的智能整合。

为了避免数据整合过程中,受到值域较大的数据条件属性值的影响,对待整合的数据样本条件属性进行归一化处理,用公式表示为:

$$a^* = \frac{a - \min a}{\max a - \min a} \quad (9)$$

式中： a^* 表示归一化的数据样本条件属性。

通过归一化处理，将多源异构网络数据样本的条件属性缩放至 $[0,1]$ 的区间内，降低大值域条件属性的影响。

采用曼哈顿距离，评估归一化后数据样本条件属性与整合规则之间的相似度，用公式表示为^[9]：

$$d(a^*(\hat{x}^i),L)=\sqrt[n]{\sum_{i=1}^n|a^*(\hat{x}^i)-L(E,N,A)|}$$

(10)

式中： d 表示曼哈顿距离； L 表示基于粗糙集挖掘的数据整合规则。

根据曼哈顿距离的计算结果，分析多源异构网络数据样本条件属性与数据智能整合规则之间的相似度。以此为基准，采用幂平均聚合策略^[10]进行数据的智能整合处理。这一过程用公式表示为：

$$H(\hat{x}^i)=\sum_{i=1}^k d(a^*(\hat{x}^i),L)^{\frac{1}{p}}$$

(11)

式中： $H(\hat{x}^i)$ 表示整合的多源异构网络数据； k 表示条件属性的匹配数量； p 表示平均聚合策略的幂。

在该策略下，将与规则相似度较高的多源异构网络数据连接，实现数据的智能整合，形成多源异构网络的数据互通。

2 算例分析

2.1 实验数据集

在多个领域所应用的数据平台中，采用增量式采集的方法集成大量的多源异构网络数据，并以原数据格式，将集成的数据存储到实验计算机中，构建实验数据集。

本次实验中，选定的多源异构网络数据集成的接口来源包括 8 种，如表 1 所示。

表 1 实验数据集集成接口来源

序号	接口来源
1	万维网联盟
2	Web 通知服务
3	传感器规划服务
4	传感器紧急服务
5	传感器观测服务
6	TML 换能器
7	SML
8	O&M

为了减小实验结果误差，将集成的实验数据随机划分为 50 个数据子集，用于开展多轮次的数据整合实践测试实验。

2.2 数据整合结果分析与讨论

在建立的实验数据集中，随机选择一个数据子集作为实验对象。采用本文方法对该数据子集中的多源异构网络数据进行智能整合处理。

通过对该数据子集内的多源异构网络数据进行元路径相似度的计算后，利用粗糙集对标准化的数据进行条件数据的分析，挖掘数据智能整合的规则。在这一过程中，计算得到数据样本的上、下近似集合。考虑到篇幅有限，仅展示其中五个数据样本的计算结果，如表 2 所示。

表 2 多源异构网络数据粗糙集条件属性约简

数据编号	上近似	下近似
1	0.75	0.45
2	0.59	0.56
3	0.91	0.45
4	0.92	0.67
5	0.59	0.52

根据如表 2 所示的上、下近似集合，挖掘该数据子集的智能整合规则。通过条件数据的相似度分析，完成数据智能整合过程。将整合结果绘制为 MDS 图的形式进行分析，结果如图 1 所示。

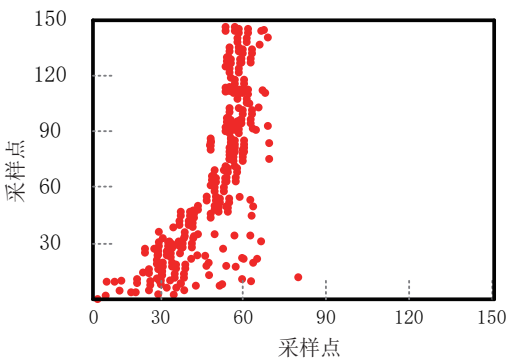


图 1 数据智能整合的 MDS 图结果

由图 1 可知，采用本文方法对多源异构网络数据进行智能整合时，整合结果表现出高度聚集的形态，且出现离群点的数量很少。此可视化结果说明，在采用文章提出基于元路径相似度与粗糙集规则挖掘的整合方法进行处理后，使得本来存在来源广泛、结构各异的网络数据可以在低维的特征空间内组成紧密并一致的簇群，准确反映出整合后数据间具备很高的内在一致性和结构相似性。

为进一步评估本文方法在数据整合时在完整性性能方面的水平，实验使用基于知识图谱和 LLM 以及图 Transformer 模型两种数据整合方法作为对比，针对 50 个数据子集展开多轮次的对比实验，并统计各方法在不同轮次实验中，数据整合效果和原始数据集间的差异，分析对应数据整合完整度。

在通过多轮次测试后，得到如图 2 所示的对比结果。分析图 2 可知，采用本文方法数据整合完整度为 98%，明显优于基于知识图谱与 LLM 的方法及基于图 Transformer 模型的方法。这一量化优势表明了模糊边界处理及冗余属性约简有

优化整合鲁棒性方面的作用,表明本文方法可以更有效地预防信息丢失的情况,适用于复杂多源异构环境下的数据整合任务。

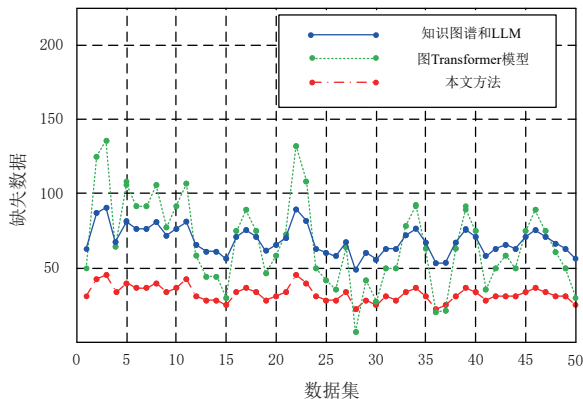


图2 数据整合结果完整度对比

3 结语

本研究设计并实现了基于粗糙集理论的多源异构网络数据智能整合方法优化研究。此方法经过引入元路径建模和PathSim相似度计算,能够识别并关联分散于不同来源及结构的同一实体数据,解决实体识别中的异构性和语义模糊性造成的影响。随后,采用粗糙集理论建立决策信息系统,经过属性约简和模糊边界的构建,可以深入挖掘数据间不确定的整合规则,明显降低冗余属性及确定分类边界所带来的信息丢失问题。再结合曼哈顿距离和幂平均聚合策略,达到高相似度匹配下数据自适应整合的目的。

实验部分在涵盖8类真实接口数据的复杂多源环境里对本方法进行系统验证。实验结果表明,采用本文方法进行整合时,数据完整度达到了98%,而丢失数据量平均只有36.8条,性能明显优于对比方法。MDS可视化分析进一步说明,整合后数据在低维空间中表现出高聚集、低离散的分布特征,表明本方法在提升数据一致性及结构完整性方面的有效性。

本研究的主要贡献在于提出一套融合元路径分析和粗糙集理论的整合框架,为处理多源异构环境下因数据来源多样、结构不一致和语义模糊带来的整合难题提供新的解决思路及方法。本方法不但具备很好的理论价值,也为实际应用中实现多源数据高质量整合与深度分析提供了可靠的技术支持。

参考文献:

- [1] 李平,马吉科,杜先波,等.融合知识图谱和LLM的不平衡样本电力数据整合[J].电子设计工程,2025,33(16):120-123.
- [2] 师瑶,段贸腾,刘丙强,等.图Transformer模型在多组学

数据整合分析中的应用[J].数学建模及其应用,2025,14(2):1-8.

- [3] 刘晓宇,金勇进.大数据背景下基于概率-非概率样本的高维数据整合推断研究[J].数理统计与管理,2025,44(3):381-392.
- [4] 宋超,刘剑磊,王志乾,等.基于大数据与聚类算法的电网信息化数据整合研究[J].信息与电脑(理论版),2024,36(20):181-184.
- [5] 陈亚颐,尹权,罗李黎,等.动态聚类下智慧油田企业多源数据整合仿真[J].计算机仿真,2025,42(3):165-169.
- [6] 赵玮,王世航.基于GIS的城市规划测绘数据整合与分析方法研究[J].城市建设理论研究(电子版),2025(20):155-157.
- [7] 宋生磊,马有录,刘晓杰,等.土地重度污染区域测绘数据整合方法研究[J].环境科学与管理,2025,50(7):126-130.
- [8] 陈亚颐,尹权,再开日亚·安尼娃尔,等.跨系统的流动性多源异构数据整合算法仿真[J].计算机仿真,2025,42(1):410-414.
- [9] 郭蕊,李奕霏,高育栋.基于人工智能技术的电力信息运维数据整合平台[J].自动化技术与应用,2025,44(6):76-79.
- [10] 张伟仁,黄旭兴.基于Mesh网络多源异构混合信息一体化融合技术研究[J].指挥控制与仿真,2025,47(2):50-53.

【作者简介】

王煥鑫(1997—),男,河南周口人,硕士,助教,研究方向:大数据、传染病动力学。

邱帅欣(1998—),女,河南濮阳人,硕士,助教,研究方向:计算机视觉、图像处理。

(收稿日期:2025-09-16 修回日期:2026-05-07)