# 利用领域本体提升数值数据集的簇质量

Sudath Rohitha Heiyanthuduwage 计算、数学与工程学院 查尔斯特大学 澳大利亚 SHeiyanthuduwage@csu.edu.au

Md Anisur Rahman 计算、数学与工程学院 查尔斯特大学 澳大利亚 ARahman@csu.edu.au

Md Zahidul Islam 计算、数学与工程学院 查尔斯特大学 澳大利亚 ZIslam@csu.edu.au

摘要— 近年来，基于本体的聚类因本体的潜在优势而受到关注。目前，基于本体的聚类方法主要应用于降低文本文档聚类中的属性维度。借助本体降低属性维度，有助于为数据集生成高质量簇。然而，基于本体的方法在数值数据集聚类中的应用尚未受到足够关注。此外，一些文献指出，基于本体的聚类可能会从数据集中生成高质量或低质量的簇。因此，本文提出一种基于领域本体的聚类方法，利用领域本体降低数值数据集的属性维度，并生成高质量簇。对于每个数据集，我们利用领域本体生成三个数据集。随后，我们使用一种名为 GenClust++ 的基于遗传算法的聚类技术对这些数据集进行聚类。我们根据误差平方和（SSE）评估各数据集的簇。我们使用六个数值数据集评估所提出的基于本体的方法的性能。实验结果表明，随着领域本体层级的提高，簇质量逐步提升。

关键词— 聚类、领域本体、聚类评估、遗传算法、数据挖掘。

## I. 引言

聚类是指将数据集中的记录划分为两个或多个簇，使相似的记录归入同一个簇，而不同的记录归入不同的簇 [1]。数据聚类的应用范围广泛，从零售数据分析到商业、天气、医疗及许多其他领域的数据分析，旨在从中识别具体且有用的数据模式[2]。通常，数据聚类针对的是记录数量和属性数量都很大的数据集。因此，聚类结果也包含大量具有众多属性的记录，若缺乏一些领域知识，通常难以解读。本体已被提出，可用于提升聚类结果，也可帮助理解聚类结果。本体被定义为特定领域的形式化规范 [3]。为特定领域定义的本体包含该领域的概念、概念的属性、概念之间的关系以及各种约束。领域本体已应用于学习、医学、工程和旅游等不同领域。特定本体中规定的领域知识有助于我们全面理解该领域。

当前基于本体的聚类方法侧重于聚类分析的不同方面，包括利用本体信息进行数据预处理、丰富词项向量、重新加权向量等。 [4]。在聚类过程中，本体可以在预处理阶段、聚类阶段或后续理解聚类结果的阶段发挥作用。例如，Hotho 等人 [5] 提出了一种利用本体进行数据预处理的技术，其算法优于其他数据预处理技术。一些方法侧重于将本体直接应用于聚类算法，并试图改进聚类过程。在 [6] 中，一种名为 Onto-SVD（奇异值分解）的算法被用于找出存储记忆之间的语义相似度。Onto-SVD 借助语义相似度和语义特征选择，结合 K 均值聚类来识别主题。

许多基于本体的聚类方法都侧重于文本文档聚类。它们分析包含大量术语的文本文件，并利用本体降低属性维度，例如 [5, 7, 8]。 [9] 设计了一种基于本体的框架，该框架应用本体，根据基于片段的部分所构成的紧密组来组织文本文件。 [10] 将基于本体的遗传算法应用于文本文件聚类。这些技术最初将文本文档中的名词表示为二维结构，或表示为文档的向量空间模型（VSM），并记录每个文档中的词频。它们通过领域专用名词而非通用术语来降低属性维度 [10]。

一些文献表明，基于本体的聚类能够提高簇质量，而另一些文献则指出，基于本体的聚类无法提高数据集的簇质量 [7]。此外，针对数值数据集的基于本体的聚类技术以及基于领域本体的技术尚未得到足够关注。因此，本文提出一种基于领域本体的聚类方法，以克服上文所述基于本体的聚类的局限性。在我们的方法中，我们使用领域本体为每个数据集生成三个数据集。随后，我们使用一种基于无参数遗传算法的

名为 GenClust 的聚类技术++ [11]对这些数据集进行聚类。我们根据误差平方和（SSE）评估每个数据集的簇。我们使用六个数值数据集来评估所提出的基于本体的聚类技术的性能。实验结果表明，随着本体层级从较低层级逐步提升至较高层级，我们的方法能够逐步提高簇质量。

本文其余部分组织如下。第 2 节介绍相关工作，第 3 节介绍我们的方法。第 4 节详述实验结果并进行讨论。第 5 节总结全文。

## II. 相关工作

数据聚类适用于多种类型的数据，包括数值数据、类别数据和文本数据。基于本体的聚类方法主要应用于文本文件聚类。文献 [12] 将基于本体的文本文档聚类所解决的问题分为四大类。1. 同义词（多个词具有相同含义）和多义现象（一个词具有多个不同含义）问题，以及使用词义消歧（WSD）方法解决这些问题。2. 术语的高维性会增加处理时间并降低簇质量。3. 从文本中提取有助于提高簇质量的核心语义特征。4. 为聚类结果提供易于区分且有意义的描述，以帮助用户理解结果。

在 [7] 中，多义名词和同义名词都被认为较为常见，并且对文档簇的形成具有根本重要性。在文档中发现同义名词时，会将它们视为本体中的同一个概念；而发现多义名词时，则将它们视为本体中的不同概念。他们的研究 [7] 表明，文档中识别出的名词本身有助于提升聚类效果。他们的研究还揭示了多义名词和同义名词在聚类中的重要性。

同义现象会降低属性维度，而多义现象则会增加属性维度 [7]。通常，文本聚类在实际应用中被认为难度较高，因为它需要在高维空间中进行聚类 [5]。Hotho 等人 [5]，借助一种基于启发式规则的算法，使用领域本体来降低属性维度。该算法也有助于识别核心语义特征。提取与领域相关的核心语义特征，有助于排除与特定领域无关的名词。识别核心语义特征不仅能进一步降低特征维度，还能为用户提供领域知识。研究发现，与使用多义名词／同义名词或使用所有名词相比，使用核心语义特征可以获得更高的簇纯度 [7]。

两种常用于基于本体的聚类的本体是 WordNet本体和基因本体（GO）。WordNet本体主要用作文档聚类方法中的参考本体，而 GO 则用作专用于医学领域的领域本体。A 基于 WordNet 的多种基于本体的聚类方法已列于 [7] 中。WordNet 也曾用于[13] ，以识别本体中各概念之间的关系。随后，根据概念及其子概念之间的关系生成了本体索引。该本体及本体索引已用于基于本体的加权聚类，以解决 Hadoop 在处理大数据时遇到的瓶颈问题 [13]。

在医学领域，聚类已应用于基因数据，以识别与特定疾病相关的特定基因。基因本体（GO）已被提出作为注释微阵列实验结果的一种方法。 [14]中提出了一种基于 R 的快速软件（csbl.go），用于高级基因注释。在 [15] 中，文档被定义为一组结构化字段，其中包括对其他数据库（如基因本体（GO））的引用，以及提供生物体相关信息的文本。查询结果以层次化文件夹的形式呈现给用户，其中包含字符串标签和簇大小。他们[15] 已应用基于本体的层次聚类，使数据库搜索更加高效有效。

Web 服务数量的增加，使用户高效、准确地找到偏好的 Web 服务变得颇具挑战性。 [16] 中提出的基于本体的聚类方法试图解决这一问题。作为解决方案，系统生成了一个基于 Web 服务相似性和特异性的本体。随后利用该本体，避免将 Web 服务错误地放入簇中，并获得更高的聚类性能。

## III. 我们提出的方法：利用领域本体提升聚类质量

本节介绍我们提出的基于领域本体的本体聚类方法。我们的方法包括六个步骤：（1）从数据集推导领域本体；（2）对数据集进行归一化；（3）从领域本体的各个层级创建用于聚类的数据集；（4）对各层级的数据集进行聚类；（5）评估聚类结果。在我们的方法中，我们重点对数值数据集进行数据聚类。我们还考虑数据集的用途，以及我们理解其应用领域的难易程度，这有助于我们为每个数据集生成领域本体。我们也检查了数据集是否包含数量适中的属性，既不太少，也不过多。我们选取的数据集包含十至三十三个属性。我们不选属性较少的数据集，是因为这样难以在领域本体中构建多个层级。如果数据集包含的属性过多，我们创建的领域本体就会有过多层级。这会导致本体之间的层级结构各不相同、无法相互比较，从而妨碍我们公平比较基于本体层级生成的数据集的簇质量。下面的小节将逐一说明我们方法的各个步骤。

### A. 步骤 1：从数据集中推导领域本体 许多现有方法都使用了 WordNet [17

] 作为参考本体。WordNet 是一个英语词汇数据库，其中包含许多通用术语，由此构成了本体中层级繁复的概念体系。我们提出一种方法，为每个数据集创建领域专用本体，以减少对通用概念的使用。创建某个数据集的领域本体时，

我们将数据集中的所有属性视为领域本体中的底层概念。随后，我们考察这些底层概念是否属于某个共同类型或种类，从而形成一个高层概念。这样便在若干底层概念与一个高层概念之间建立了包含关系。 [18] 包含关系可以视为父子关系、类型关系或种类关系。我们采用启发式方法来识别高层（父级）概念。在下面的小节中，我们将详细说明如何为 Travel Review (TR) 数据集创建领域本体，并介绍这一过程。随后，我们也将同样的流程应用于其他数据集，为全部六个数据集创建领域专用本体。

表 I。 本体的初始属性和一级概念

| # | 属性 | # | 属性 | # | 属性 |
|---|---|---|---|---|---|
| 1 | 教堂 | 9 | 餐厅 | 17 | 个游泳池 |
| 2 | 度假村 | 10 | 个酒吧 | 18 | 个健身房 |
| 3 | 海滩 | 11 | 个本地服务 | 19 | 家面包店 |
| 4 | 公园 | 12 | 家汉堡/披萨店 | 20 | 家美容院和水疗中心 |
| 5 | 剧院 | 13 | 酒店 & 其他住宿设施 | 21 | 家咖啡馆 |
| 6 | 博物馆 | 14 | 家果汁吧 | 22 | 个观景点 |
| 7 | Mall | 15 | 家美术馆 | 23 | 纪念碑 |
| 8 | Zoos | 16 | 家舞蹈俱乐部 | 24 | 座花园 |

1）识别领域本体的一级概念 这是我们创建领域本体过程中遵循的第一步，我们按照以下方式进行创建

领域本体。除序列号、所有记录取值相同的属性以及类别属性（如果存在）之外的所有属性，都被视为本体的一级概念。例如，TR 数据集包含 25 个属性，但第一个属性用户 ID 是序列号。因此，除该属性外的所有其他属性均列于表 I 中。这 24 个属性构成了名为旅游评论本体（TRO）的领域本体的一级概念，该本体用于描述 TR 领域的领域知识。表 I 中的这 24 个属性将在下一步中用于构成二级概念，即 TRO 的下一个层级。

2）推导本体的二级概念 在这里，我们可以看到哪些底层概念属于

将其归入哪些共同的高层概念，并将这些高层概念作为 TRO 的二级概念。高层概念基于本体工程师对领域的理解进行定义。我们认为，Churches、Museums、Art Galleries 和 Monuments 这些概念属于高层概念 Cultural Place（表 II）。Beaches 和 Viewpoints 这些概念属于高层概念 Natural Place；Parks、Malls、Zoo 和 Gardens 这些概念属于高层概念 Natural Place。Parks、Malls、Zoo 和 Gardens 这些概念属于高层概念 Social Place。

表 II。 从一级概念中识别出的二级概念

| # | 二级概念 | 一级概念 |
|---|---|---|
| 1 | 文化场所 | 教堂，博物馆，Art 画廊，纪念碑 |
| 2 | 自然场所 | 海滩，观景点 |
| 3 | 社交场所 | 公园、购物中心、动物园、花园 |
| 4 | 住宿 | 度假村，酒店和其他住宿设施 |
| 5 | 娱乐 | 剧院、舞蹈俱乐部 |
| 6 | 餐饮 | 餐厅、酒馆和酒吧，汉堡与披萨店、果汁吧，面包店、咖啡馆 |
| 7 | 服务 | 本地服务，游泳池、健身房、美容与水疗 |

高层概念 Accommodation 包含低层概念 Resorts、Hotels 和 Other Lodgings。高层概念 Entertainment 包含低层概念 Theatres 和 Dance Clubs。高层概念 Food and Drinks 包含低层概念 Restaurants、Pubs and Bars、Burger and Pizza Shops、Juice Bar、Bakeries 和 Cafes。高层概念 Services 包含低层概念 Local Services、Swimming Pools、Gyms、Beauty and Spa，如表 II 所示。下一步，我们使用这七个二级概念来识别 TRO 的三级概念。

3) 再次推导本体的第 3 级概念：为第 2 级确定的七个概念 o

TRO 用于形成两个第 3 级高层概念：景点和设施。第 2 级概念文化场所、自然场所和社交场所属于高层概念景点。第 2 级概念住宿、娱乐、餐饮和服务属于高层概念设施，如表 III 所示。现在，TRO 的第 3 级仅有两个概念，无法再向上构建更高层级。这两个第 3 级概念现被视为本体根概念“事物”的子概念。尽管我们在这些数据集的领域本体中仅得到 3 层概念层次结构，但本体工程师在为其他数据集构建领域本体时，可能会得到不同数量的层级。

TABLE III. 根据第 2 级概念识别出的第 3 级概念 概念

| # | 第 3 级概念 | 第 2 级概念 |
|---|---|---|
| 1 | 景点 | 文化场所、自然场所、社交场所 |
| 2 | 设施 | 住宿、娱乐、餐饮和饮品、服务 |

4）旅游评论本体 当这些包含关系

考虑所有三个层级概念之间的关系后，我们得到了 TR 数据集的领域本体。我们使用本体编辑工具 Protégé 创建的 TRO 本体 [19] 见图 1。其中，owl:Thing 是根概念，根概念中的前缀 owl: 指的是网络本体语言（OWL）[20]。

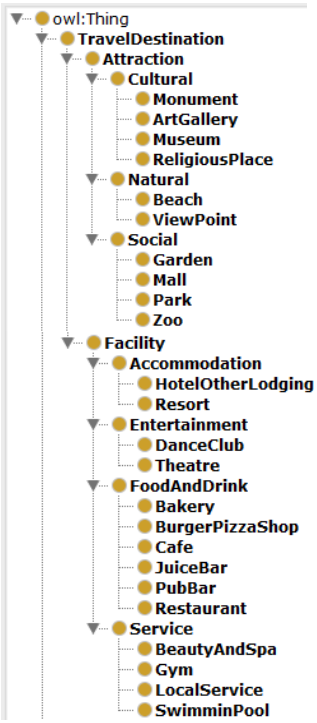


图 1. TR 数据集的领域本体。

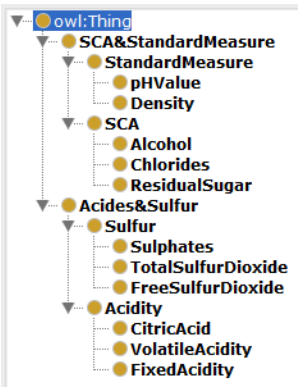


图 2. WQ 数据集的领域本体。

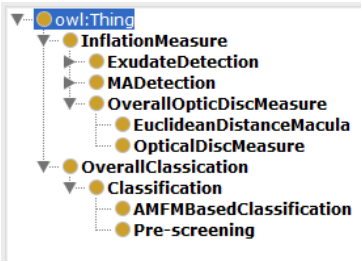


图 3. DRD 数据集的领域本体。

1) 每个数据集的领域本体 我们采用了相同的方法

并重复步骤 3.1.1 至 3.1.5，以推导其他数据集的领域本体。例如，图 2 和图 3 展示了为葡萄酒质量（WQ）和德布勒森糖尿病视网膜病变（DRD）数据集构建的领域本体。每个 6 本体包含三个层级的包含层次结构。因此，所有六个领域本体都可以相互比较，尤其是在用于对数据集进行聚类时。

B. 第 2 步：规范化数据集

某些数据集需要进行数据归一化，以简化数据，使其适用于聚类，并使聚类结果具有可比性。本研究使用的所有原始数据集均为数值型数据集。不过，这些数据集中的各个属性的取值范围各不相同，有些值小于 0，有些值大于 100。因此，我们对每个数据集中的属性值进行归一化，使所有值都介于 0 和 1 之间。一种简单的属性值归一化方法是将其除以该属性的最大值。然而，这种方法的缺点是，如果数据集中包含取值为负数的属性，某些归一化值就会变成负数。为解决这一问题，我们使用了一个特定公式，以确保所有归一化值都介于 0 和 1 之间 [21]。

C. 步骤 3：从领域本体的各个层级创建用于聚类的数据集

在此步骤中，我们针对 3.1 节中创建的每个领域本体，逐一为本体的每个层级创建数据集。为清楚说明这一子任务，以下描述以 TRO 为例。TR 数据集的领域本体在包含层次结构中包含三个层级。TRO 最低层级的 24 个概念对应原始数据集的 24 个属性。从第 1 级向上到第 3 级时，我们将低层级属性概括为高层级属性，因此高层级的属性数量会减少。由于 TRO 的第 1 级有 24 个概念，我们将原始数据集的 24 个属性作为第 1 级数据集。TRO 的第 2 级有 7 个概念，因此我们将 TRO 第 1 级的数据集转换为包含 7 个属性的数据集。为此，我们计算每组第 1 级属性的平均值，并将其作为领域本体中对应第 2 级概念的第 2 级属性值。这里我们使用属性的平均值，以便每个子属性具有相同权重，从而对聚类过程产生同等影响。TRO 在最高层级（第 3 级）有两个属性。同样，我们计算每组第 2 级属性值的平均值，并将其作为第 3 级的属性值。

D. 步骤 4：对数据集进行聚类

准备好每个本体三个层级的数据集后，我们使用聚类算法 GenClust++ [11] 对全部 6 个数据集的 3 个层级进行聚类。GenClust++ 被提出作为一种增强型聚类算法，与当前大多数可用的聚类算法相比，它能够生成更好的聚类结果。GenClust++ 可在数据挖掘环境 WEKA [22] 中使用，全部 18 个数据集的聚类均在 WEKA 中完成。由于 GenClust++ 是一种无参数聚类算法，我们无需指定簇数，它会自动确定数据集的簇数。

E. 步骤 5：评估聚类结果

对每个数据集的聚类结果进行评估，以判断聚类质量是否有所提升。为此，

我们首先找出由领域本体最低层级（第 1 级）产生的每个簇中的记录。然后，将较高层级（第 2 级和第 3 级）数据集的聚类结果，与第 1 级数据集的聚类结果中相同的记录进行比较。本研究使用一种名为误差平方和（SEE）的聚类评估标准来评估聚类结果。[1, 21]计算本体不同层级下各数据集所产生的每个簇的 SEE 值。此项计算基于从全部六个原始数据集派生的数据集进行。

## IV. 实验结果与讨论

为评估我们的基于本体的聚类方法的性能，我们从 UCI 机器学习资料库中选取了六个数值数据集 [23]。表IV展示了所选数据集的概要，包括每个数据集的名称、记录数和属性数。在这些数据集中，玻璃识别（GI）数据集的属性数和记录数最少。土耳其学生评估（TKS）数据集的属性数和记录数最多。

表IV。 所选数据集概要

| # | Name | # of 记录 | # of 属性 |
|---|---|---|---|
| 1 | 德布勒森糖尿病视网膜病变 (DRD) | 1151 | 20 |
| 2 | 葡萄酒质量（WQ） | 4898 | 12 |
| 3 | 玻璃识别（GI） | 214 | 10 |
| 4 | 土耳其学生评估（TKS） | 5820 | 33 |
| 5 | 图像分割（IS） | 2310 | 19 |
| 6 | Travel Review（TR） | 5456 | 25 |

### A. 实验结果 我们进行评估

六个数据集的聚类结果，以评估我们的基于本体的聚类方法。我们考察领域本体从最低层级到最高层级的过程中，聚类质量是否得到提升。评估分别在 6 个数据集上独立进行。我们在这六个数据集上获得的聚类结果使用 SSE 进行评估。根据这一方法，如果簇的 SSE 值较低，则认为该簇质量较好。六个数据集 DRD、WQ、GI、TKS、IS 和 TR 在 3 个层级上的 SSE 值分别显示在表 V 的各列中。该表的每一行包含 6 个领域本体中某一特定层级的 SEE 值。

表 V. 六个数据集中 SSE 的改进情况（越低越好）。

| L | DRD | WQ | GI | TSK | IS | TR |
|---|---|---|---|---|---|---|
| L1 | 1775.73 | 3646.70 | 217.26 | 123524.09 | 4430.54 | 43877.41 |
| L2 | 1342.16 | 3287.27 | 217.26 | 132254.34 | 4424.24 | 77332.15 |
| **L3** | **830.83** | 2451.24 | 176.36 | **121357.81** | 4419.81 | 73989.66 |

接下来，我们根据 SEE 值从最低层级到最高层级的变化，计算聚类质量的提升百分比。表 VI 将其列为两项连续的改进幅度以及 SSE 值的总改进幅度。也就是从第 1 级到第 2 级，再从第 2 级到第 3 级的改进。总提升百分比是前两项改进幅度的汇总。

表 VI. 六个数据集中的误差平方和改进百分比（越低越好）。

| 增强 | DRD | WQ | GI | TSK | IS | TR |
|---|---|---|---|---|---|---|
| L1 到 L2 | 24.42 | 9.86 | 0 | -7.07 | 0.14 | -76.25 |
| 第 2 级到第 3 级 | 38.10 | 25.43 | 18.83 | 8.24 | 0.10 | 4.32 |
| 总计 | 62.51 | 35.29 | 18.83 | 1.17 | 0.24 | -71.92 |

在表 VI 中，DRD 数据集的误差平方和（SSE）值改善幅度最高（62.51）。WQ 数据集的改善幅度位居第二，为 35.29%；GI 数据集的改善幅度为 18.83%，位列 3rd 最高。TSK 和 IS 数据集的改善幅度较小，分别为 1.17 和 0.24。TSK 数据集的整体表现略有改善，且第 2 级到第 3 级有所改善，但第 1 级到第 2 级的聚类质量有所下降。尽管最后一个数据集 TR 的 SEE 总体上没有改善，但其第 2 级到第 3 级改善了 4.32%。所有数据集在至少两个层级的聚类质量都有所改善。我们观察到，TR 数据集的 SEE 负值相对较高，出现在

从第 1 级到第 2 级。造成这一偏差的原因是 GenClust++在第 1 级产生了较多簇，共 71 个。我们使用 GenClust++ 对 6 个数据集的 3 个层级进行聚类实验，结果见表 VII。尽管我们可以为不同数据集得到不同的簇数，但对于大多数数据集，GenClust++ 在每个层级都会生成 2 到 4 个簇。

表 VII。 GENCLUST++ 针对本体各层级生成的簇数

| 本体层级 本体 | DRD | WQ | GI | TSK | IS | TR |
|---|---|---|---|---|---|---|
| L1 | 4 | 2 | 2 | 3 | 2 | 71 |
| L2 | 4 | 2 | 2 | 2 | 2 | 2 |
| L3 | 30 | 40 | 2 | 3 | 2 | 3 |

### B. 讨论

在我们的工作中，我们研究了聚类质量如何随本体层级的变化而变化。为此，我们为从 UCI 机器学习资料库获取的 6 个数值数据集创建了领域特定本体。根据我们开展的聚类实验和评估，可以看出，从领域本体的最低层级到最高层级，簇质量有了显著提升。

许多现有的基于本体的聚类技术（包括[7, 12] ）使用了本质上具有通用性的 WordNet 作为参考本体。WordNet本体的包含层级也包含许多层级。他们的工作中使用的领域本体包含物质事物、特殊概念和无形事物等过于宽泛的概念。纳入通用概念会导致许多或全部记录被归入一个

单个簇，这会妨碍识别具有领域特定特征的簇。因此，在一些研究中，聚合低维概念空间时的评估结果并未达到预期[5]。

在我们的工作中，我们提出根据数值数据集的属性创建领域特定本体。由于这些属性能够具体描述数据集中的值，因此可以将其视为领域特定概念。这有助于我们避免在聚类中计入非领域特定的通用概念，从而缓解由通用概念导致的上述簇质量较差问题。通过忽略出现频次过高和过低的概念，可以降低用于聚类的属性维度 [5]。然而，这些属性对于聚类而言可能更为重要。在我们的方法中，我们对一组较低层级的（子）概念取平均值，并通过沿本体的包含层级向上移动来引入较高层级的（父）概念。我们所使用的这些较高层级概念是领域特定的概念，也是该领域具有代表性的概念。在他们的方法中 [5] ，他们采用自顶向下的方法从本体中选择概念。本工作中使用的算法 [5] 忽略了一些可能有助于提升聚类质量的概念。然而，在我们的方法中，我们采用自底向上的方法。

## V. 结论

本文提出了一种基于领域本体的聚类方法。我们的目标是将领域本体融入聚类，以提高数值数据集的簇质量。我们根据数值数据集的属性构建了一个领域本体。在我们的方法中，每个本体包含三个层级，因此我们从单个本体生成了三个数据集。随后，我们使用一种称为 GenClust++ 的无参数遗传算法聚类技术对这些数据集进行聚类。我们使用误差平方和（SSE）对簇进行评估。我们方法的 SSE 值表明，对于大多数数据集而言，随着本体层级从较低层级提升至较高层级，聚类质量逐步提高。六个数据集中的前五个数据集的簇质量分别提高了 62.51%、35.29%、18.83%、1.17% 和 0.24%。

我们认为，通过更深入地理解领域并构建高度领域特定的本体，可以进一步提高簇质量。引入领域专家并对该领域进行深入研究，将有助于实现这一点，并对聚类结果产生积极影响。未来，我们计划扩展这项工作，开展进一步评估。此外，我们还计划使用属性数量更多的数据集对该方法进行评估，这些数据集将构建出具有超过 3 个包含层级的领域本体。我们还计划使用不同的聚类算法和不同的评估标准，对该方法的簇质量进行评估。

## 参考文献


[1] A. K. Jain, 《数据聚类：超越 K-means 的 50 年》, 《模式识别快报》, 31(8), 651-666, 2010。

[2] Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agus [illegible] 最先 [illegible] 《人工智能工程应用》, 110, 104743。[3] T. R. Gr ıbe [illegible] 于用于知识共享的本体设计原则》, 《国际人机研究杂志》, 43(5-6), 907-928, 1995。[4] A. Hotho, A. Maedche and S. Staab, 《基于本体的文本文档聚类》, KI, 16(4), 48-54, 2002。[5] L. Shi and R. Setchi, 《面向用户的存储记忆基于本体的聚类》, 《应用专家系统》, 39(10), 9730-9742, 2012。[6] S. Fodeh, B. Punch and P. N. Tan, 《基于核心语义特征的本体驱动文档聚类》, 《知识与信息系统》, 28(2), pp. 395-421, 2011。[7] A. Valls, K. Gibert, A. Orellana and S. Antón-Clavé, 《利用基于本体的聚类理解英国游客造访地中海沿岸目的地的推动与吸引因素》, 《信息与管理》, 55(2), 145-159, 2018[8] A. Tagarelli and G. Karypis, 《一种面向多主题文档的分段式聚类方法》, 《知识与信息系统》, 34(3), 563-595, 2013。[9] W. Song, C. H. Li, and S.C. Park, 《利用本体进行文本聚类并评估各种语义相似度度量有效性的遗传算法》, 《应用专家系统》, 36(5), 9095-9104. 2009[10] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, 《一种利用 WordNet 和词汇链进行文本聚类的语义方法》, 《应用专家系统》, 42(4), 2264-2275. 2015。[11] K. Ovaska, M. Laakso, and S. Hautaniemi, 《用于微阵列实验的快速基因本体聚类》, 《生物数据挖掘》, 1(1), 11, 2008。[12] C. Corsi, P. Ferragina and R. Marangoni, 《BioPrompt-box：一种用于在生物数据库中搜索的基于本体的聚类工具》, 《BMC 生物信息学》, 8(1), S8, 2007。[13] D. Pradeep, and C. Sundar, 《QAOC：用于 Hadoop 中数据管理的新型查询分析与基于本体的聚类》, 《未来一代计算机系统》, 108, 849-860, 2020。[14] D. Dua and C. Graff, UCI 机器学习资料库[http://archive.ics.uci.edu/ml]。加利福尼亚州尔湾：加利福尼亚大学信息与计算机科学学院，2019。[15] G. A. M [illegible] 39-41 [illegible] 《通过特异性感知的本体聚类缓解稀疏性，以改进 Web 服务推荐》, 《日本电气学会电气与电子工程汇刊》, 14(10), 1507-1517, 2019。[17] V. C. Storey, 《理解语义关系》, 《VLDB 杂志》, 2(4), 455-488. 1993。[18] M. A. Musen, 《Protégé 项目：回顾过去，展望未来》, 《AI Matters》, 1(4), 4, 2015。[19] D. L. McGuinness and F. Van Harmelen, 《OWL Web 本体语 [illegible] 《无 [illegible] 成高质量簇的高质量初始种子自动选择》, 博士论文，查尔斯特大 [illegible] 1。[21] M. Z. Islam, V. Estivill-Castro, M. A. Rahman and T. Bossomaier, 《通过新颖的遗传算子编排结合 K 均值聚类与遗传算法，实现高质量聚类》, 《应用专家系统》, 91, 402-417, 2018。[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, 《WEKA 数据挖掘软件：更新》, 《ACM SIGKDD 探索通讯》, 11(1), 10-18, 2009。[23] D. Dou and H. L. H. Wang, 《语义数据挖掘：基于本体的方法综述》, 载于：IEEE 国际语义计算会议 (ICSC)，2015。