arXiv:2305.06528v1 [cs.DB] 11 May 2023

一种用于植被数据集成的半自动化混合模式匹配框架

MdAsif-Ur-Rahman[a] (rahman.m.asifur@gmail.com), Bayzid Ashik Hossain[a](bayzid.hossain@det.nsw.edu.au), Michael Bewong[b] (mbewong@csu.edu.au), Md ZahidulIslam[a] (zislam@csu.edu.au), Yanchang Zhao[c](yanchang.zhao@data61.csiro.au), Jeremy Groves[d] (jeremy.groves@cer.gov.au), Rory Judith[d] (rory.judith@environment.gov.au)

[a] 查尔斯特大学计算、数学与工程学院，Panorama Avenue，巴瑟斯特，新南威尔士州 2795，澳大利亚。[b] 查尔斯特大学计算、数学与工程学院，Boorooma street，邮政信箱 588，沃加沃加，新南威尔士州 2678，澳大利亚。[c] CSIRO Data61，GPO Box 1700，堪培拉，澳大利亚首都领地 2601，澳大利亚。[d] 气候变化、能源、环境与水务部，GPO Box 3090，堪培拉，澳大利亚首都领地 2601，澳大利亚。

通讯作者：

Michael Bewong，计算机、数学与工程学院，查尔斯特大学，Boorooma街，邮政信箱588号，澳大利亚新南威尔士州瓦加瓦加市，邮编2678。电话：+61 02 6933 2591电子邮箱：mbewong@csu.edu.au

摘要

整合不同且分散的植被数据，对于制定一致且有充分依据的国家政策和开展管理工作至关重要。澳大利亚气候变化、能源、环境与水务部（DCCEEW）下属的全国植被信息系统（NVIS），是唯一一个全国一致的植被数据库，并提供了不同地理位置植被类型的分层类型学。目前，该数据库采用人工方法整合来自各州和领地的不同数据集，这种方法不仅耗费大量人力，而且容易出现人为错误。为应对从异构数据源获取最新植被数据的需求日益增长这一挑战，本文提出了一种半自动混合匹配器（SAHM）。SAHM采用双层匹配框架，同时利用模式级匹配和实例级匹配。本文还提出了SAHM中的一项关键创新技术——多变量统计匹配，用于自动化模式评分。该技术充分利用领域知识以及属性之间的相关性，以提高匹配效果。为验证所提出框架的有效性，本文评估了SAHM各个组成部分单独使用及组合使用时的性能。实证评估结果表明，所提出的框架具有良好的有效性，其性能优于现有最先进的方法，包括Cupid、Coma、相似性泛洪、Jaccard-Leven匹配器、基于分布的匹配器和EmbDI。具体而言，与最先进的技术相比，SAHM的准确率达到88%至100%，F1分数也显著更高。此外，研究还表明，SAHM的效率比现有技术高出多个数量级。

关键词：数据集成、模式匹配、模式映射

1. 引言

国家植被信息系统（NVIS）（NVIS 技术工作组，2017）是澳大利亚各方持续开展的一项协作倡议。

联邦政府以及各州和领地政府。它是唯一一种基于植被结构数据与植物区系[1]数据之间的关系，对植被信息进行全面描述和表示的全国性综合手段。NVIS 持续更新，以改进澳大利亚境内的植被规划和管理，例如应对丛林火灾等事件。它还为众多国家和国际层面的法律、政策及报告提供了基础。然而，NVIS 面临的主要挑战之一是如何将异构数据集整合到联邦数据库中。

联邦政府强烈推动以数据为依据制定政策（Thackway et al., 2007; Haslem et al., 2010; Klein et al., 2009）。为提供这些洞见，联邦政府需要快速整合 NVIS 中彼此不同的数据集。随着许多新数据源不断涌现，部门对数据的需求也在迅速增长。遗憾的是，模式匹配本质上仍是一项手动且耗费人力的工作，通常需要较高水平的领域知识。所需工作量通常随待匹配的模式数量线性增长，同时复杂度也不断增加（Smith et al., 2009）。这限制了流入该部门的数据，因为整合新数据源往往需要投入大量资源。自动化模式匹配将显著缓解该部门数据汇集工作的这一瓶颈，并扩充部门的知识库。

模式匹配研究已经持续了三十多年。目前，研究人员正转向模式匹配自动化，以识别数据库属性之间的对应关系（Dongming & Guohua, 2010; Gal et al., 2019）。现有技术通常采用均值和标准差等简单统计方法（ Aumueller et al., 2005; Coen & Xue, 2009; Chaudhuri et al., 2006），以及基于字符串的比较方法，例如前缀/后缀测试和编辑距离（Zhang 等人，2011a; Rahm & Bernstein, 2001; Kang & Naughton, 2003; Shrestha 等人，2019）。然而，这些方法面临的挑战在于，表示相同概念的异构模式……

[1]对某一区域内存在的植物物种进行科学研究。

可能具有不同的结构和命名格式。同时，不同的模式可能具有相同的结构和命名格式，但所表示的信息内容却截然不同。此外，在数据集成过程中，某些属性类型（例如数值型）可能需要映射为不同的属性类型（例如类别型）。遗憾的是，现有基于模式或实例的匹配技术无法充分应对这些挑战。

本文提出了一种新颖的混合模式匹配框架，用于实现 NVIS 数据库集成的自动化。本研究由气候变化、能源、环境与水资源部（DCCEEW）[2],和 CSIRO 的 Data61 资助。为有效实现整个流程的自动化，我们提出了 SAHM，这是一种采用模式层匹配和实例级匹配的双层解决方案，旨在应对上述挑战。第一层称为模式层匹配器，依赖模式层信息，由两个匹配器组成，即领域知识匹配和语言匹配。第二层称为 实例级匹配器，依赖实例级信息，由两个匹配器组成，即单变量统计匹配和多变量统计匹配。在所提出的框架中，领域知识匹配通过一组预定义规则融入领域专家的知识；语言匹配采用基于属性的相似性度量；单变量统计匹配利用属性内的值分布；多变量统计匹配则利用多个属性之间的相关性，以确保实现更准确的模式匹配。

尽管实证部分证明，各种匹配技术都显著提升了 SAHM 的有效性，但 SAHM 的核心在于新颖的多变量统计匹配。该技术以源数据集和目标数据集中任何已知的匹配对为基础。也就是说，当用户从标记为 a^s 的源数据集中选择一个属性（例如州数据集），然后从标记为 a^d 的目标数据集中选择另一个已知与之匹配的属性（例如 NVIS 数据库）时，所提出的新颖方法...

[2]原农业、水务与环境部（DAWE）

该技术计算源数据集（分别为目标数据集）中所选属性与其余属性之间的相关性。随后，该技术根据其余属性相对于所选源属性（分别为目标属性）的相关性得分，尝试寻找潜在的源属性与目标属性匹配对。我们假设，如果某个源属性与目标属性相匹配，那么它们分别与a^s 和 a^d 之间的相关性也应相似。因此，利用这种相关性，多变量统计匹配方案能够有效识别源属性与目标属性之间的匹配关系。

本文其余部分组织如下。第 2 节介绍这一模式匹配挑战的背景，并综述模式匹配领域的相关文献。第 3 节介绍本文提出的解决方案，将一些最先进的模式匹配技术，如 Levenshtein 相似度（Levenshtein et al., 1966）、Monge-Elkan 相似度（Monge et al., 1996）、TF-IDF（Salton 和 McGill, 1983）以及单变量统计匹配，与我们提出的新型多变量统计匹配技术巧妙地结合起来。第 4 节通过分析各个匹配器的性能以及 SAHM 整体框架的表现，展示所提出框架的有效性。最后，第 5 节给出结论。

2. 背景与相关工作

在本节中，我们介绍国家植被信息系统（NVIS）的模式匹配环境。接下来，我们讨论模式映射领域的相关文献，并分析将现有技术应用于独特的 NVIS 模式映射环境时所面临的挑战。表 1 总结了本文中常用的符号。

2.1. NVIS 模式映射环境

国家植被信息系统（NVIS）（NVIS 技术工作组，2017）是澳大利亚

表 1：符号说明

符号	含义
R^{s_i}、 a^{s_i}、n_{s_i}	源数据、源属性（状态 i）、 R^{s_i} 中的属性数量
R^d、 a^d、 n_d	目标数据、目标属性（NVIS）， R^d 中的属性数量
M^{sl}, $score^{sl}$	模式层匹配器、模式级匹配得分
M^{il}, $score^{il}$	实例级匹配器、实例级匹配得分
pc	皮尔逊相关系数得分
Sim_*	相似度函数 *

联邦政府与各州及领地政府持续开展的一项合作计划。下文中，我们将州政府和领地政府统称为州政府。NVIS 将各州政府提供的结构和植物区系等植被信息整合到联邦统一数据库中。图1展示了这一情况。图中，NVIS 数据（又称目标数据）表示为 $R^d(a_1^d, \cdots, a_{n_d}^d)$。各州数据（又称源数据）表示为 $R^{s_i}(a_1^{s_i}, \cdots, a_{n_{s_i}}^{s_i})$，其中 i 表示某个特定州的实例。NVIS 框架旨在构建一个综合数据系统，汇集不同州的植被类型范围及其分布信息。该框架包含两个数据库，分别表示两个抽象层级，即 *NVIS* 详情 （ND）和*NVIS* 扁平化 （NF）。ND 提供每条记录的详细信息，其属性包含有关层次、物种和结构的信息。NF 则是 ND 的一种便捷表示，其中 NF 中的每个属性（或属性值）均由 ND 中两个或更多属性（或属性值）组合而成。图2展示了这两个数据库。例如，在当前的 NVIS 方案中，一个 NF 属性值为“$U+\hat{}Acaciasclerosperma, Acaciavictoriae\backslash\hat{}tree\backslash 6\backslash i; M\hat{}Acaciaeremaea, \hat{} Acaciaramulosavar.linophylla, Sennasp.\backslash\hat{}shrub\backslash 4\backslash r; GCenchrusciliaris\backslash\hat{} tussockgrass\backslash 2\backslash r$”，其中以“;”分隔的部分分别表示层次、物种和结构信息。每个值代表一条独立的信息

例如 “U+” 即代表 “上层植被层 ” 在 NF 属性值中

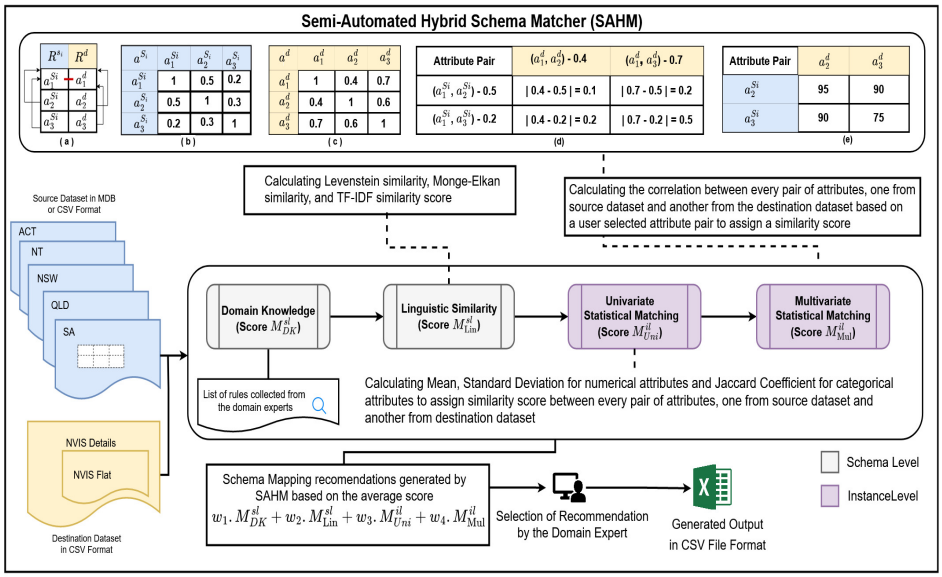


图1：所提出方法的系统架构

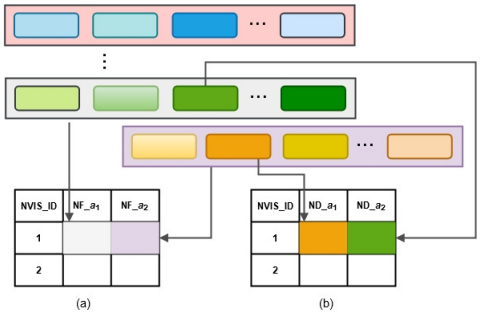


图2：NVIS 数据库的不同方案 (a) NVIS 扁平（NF）方案 (b) NVIS 详情（ND）方案

上文表示分层信息。然而，在 ND 方案中，这些信息将作为单独的属性值记录。

各州都维护自己的植被数据库，其模式可能与 NVIS 模式不同。此外，这些数据库通常以不同格式存储，例如，澳大利亚首都领地（ACT）、北领地（NT）、西澳大利亚州（WA）和南澳大利亚州（SA）

使用 CSV 格式，而昆士兰州（QLD）使用 MDB 格式。通常的做法是在将数据集成到 NVIS 框架之前，手动将所有不同格式转换为同一种格式。目前，将州数据集集成到 NVIS 框架时，领域专家会遵循《澳大利亚植被属性手册》（第 7 版）（AVAM）（NVIS 技术工作组，2017）中规定的数据映射规则。例如，《AVAM》表 7 所述，植被覆盖度的取值 70-100、30-70、10-30、 <10、0 和 0-5，可分别表示为取值为“d”、“c”、“i”、“r”、“bi”和“bc”的覆盖代码。因此，AVAM 在人工数据集成过程中通过提供指导方针发挥着至关重要的作用。然而，植被调查实践中的不一致使这一过程变得复杂，可能导致数据看似更符合 ND 或 NF 模式之一，但实际上并未完全匹配。这一集成过程面临的主要挑战是确定哪个 NVIS 模式与州数据最匹配，并进一步准确识别各数据库中相匹配的属性。

2.2. 相关工作

数据集成领域已经开发了多种模式匹配技术（Bilke 和 Naumann，2005；Zhang 等人，2010、2011b；Fernandez 等人，2018；Cappuzzo 等人，2021）和工具（Aumueller 等人，2005；Madhavan 等人，2001；Do 和 Rahm，2002；Seligman 等人，2010；Chen 等人，2018；Hu 等人，2008；Cruz 等人，2009）。CUPID（Madhavan 等人，2001）是一种混合方法，旨在改进通用模式匹配；它利用模式元素、模式结构和基于语言的匹配来映射模式。COMA（Do 和 Rahm，2002）是一种通用模式匹配方法，能够以灵活的方式结合简单匹配器、混合匹配器和面向复用的匹配器。COMA 也提供了适宜的环境，用于测试不同算法并评估其单独或组合后的性能。后来，他们提出了增强版本 COMA++ （Aumueller 等人，2005；Do 和 Rahm，2007；Melnik 等人，2002），该版本支持大量匹配器，并要求用户干预以选择最佳的匹配器组合。然而，COMA++ 复用预-

之前的匹配结果要求新数据集具有相同的领域和实体。此外，如果用户缺乏对系统的适当了解，匹配器选择这一操作的可行性便会受到质疑。DUMAS（Bilke 和 Nau- mann，2005）是一种基于实例的模式匹配算法，它利用数据集中的重复项进行匹配。基于分布的匹配器（Zhang 等人，2011b）也是一种基于实例的方法，通过比较不同列各自的数据值来捕捉它们之间的关系。基于分布的匹配器使用列对之间的地球移动距离（EMD）来实现这一目的。为了自动化蛋白质结晶分析，Shrestha 等人（2019）提出了一种基于语言的模式匹配技术，根据相似度值匹配模式元素。然而，仅考虑语言相似性无法分析元素级属性。EmbDI（Cappuzzo 等人，2021）是一种通过构建关系嵌入来促进关系数据模式匹配的框架。EmbDI 通过评估两个数据集对应的嵌入，找出它们的列之间的关系。

基于各种模式匹配技术，已有大量工具被开发出来。Harmony 匹配引擎（Seligman 等人，2010）是一种半自动化模式匹配工具，它在传统模式匹配技术中加入了对文本文档进行语言学预处理的方法，以加快发现两个模式之间相似性的过程。为查看和修改识别出的相似模式，该匹配器采用基于评分的策略，将多种匹配算法与图形用户界面相结合。FlexMatcher 是另一种开源工具（Chen 等人，2018），它使用一系列监督式机器学习技术来训练模式匹配器。此外，Falcon-AO（Hu 等人，2008）和 AgreementMaker（Cruz 等人，2009）也是其他一些支持结合使用不同语言学、结构和实例匹配方法的模式匹配工具。

2.3. 挑战与机遇

将现有技术直接应用于我们独特的 NVIS 场景时，主要存在 3 个问题。首先，现有技术未提供将领域知识纳入模式匹配过程的方法。事实上，现有技术通常经过泛化处理，且不依赖特定领域（Shrestha 等人，2019）。在我们的 NVIS 场景中，我们发现，可以利用现有领域知识改进模式匹配任务，同时不增加匹配方案对领域知识是否存在的依赖。

其次，现有的模式匹配与映射技术（Madhavan 等人，2001；Do 和 Rahm，2002；Zhang 等人，2011b；Fernandez 等人，2018；Cappuzzo 等人，2021）尚未充分探索利用源数据集 R^s 中属性之间的相关性，将其属性映射到目标数据集 R^d中的属性。我们将这种新方法称为多变量统计匹配。该方法的核心基于以下假设：若已确定任意两个属性 a^s 和 a^d 相匹配，那么如果属性 a^s_j 与 a^s 的相关性最高，且 a^d_k 与 a^d的相关性最高，则 a^s_j 和 a^d_k 很可能等价。

第三，许多现有系统要求输入为关系数据库和本体（Zhang 等人，2011a；Madhavan 等人，2001）。然而，实际情况并非总是如此。如前所述，NVIS 数据集可能并不严格符合关系模型的基本原则（参见 NF）。与此同时，各州并未为其数据集模式提供任何本体。因此，任何拟议的解决方案都必须考虑到这一点。

接下来，我们将介绍所提出的解决方案如何利用上述机会，应对 NVIS 场景下的模式映射挑战。

3. 提出的方法

在本研究中，我们提出了一种称为$SAHM$ 的双层集成模式匹配决策支持框架，如图1所示。第一层的功能

在 SAHM 中，称为 模式层匹配器 并表示为 $M^{sl}(R^d, R^s) = score^{sl}$的函数分别从 R^d 和 R^s 中提取模式信息，并返回针对每个属性对 $\langle a^s, a^d \rangle$的匹配分数列表 $score^{sl}$ 。类似地，第二层函数称为 实例层匹配器 ，表示为 $M^{il}(R^d, R^s) = score^{il}$，分别从 R^d 和 R^s 中的实例提取信息，并返回针对每个属性对 $\langle a^s, a^d \rangle$的匹配分数列表 $score^{il}$ 。最后，将两个分数 $score^{sl}$ 和 $score^{il}$ 进行聚合，为每个属性对 $\langle a^s, a^d \rangle$生成一个最终分数，记为 $score$。因此，SAHM 以两个数据库作为输入，并生成一个元组集合，其中包含所有属性对及其各自的分数即 $SAHM(R^d, R^s) = \{(\langle a^s, a^d \rangle, score)_i\}$。

我们注意到，每个匹配函数 M^{sl}(分别为 M^{il}) 都是由多个匹配函数组成的集成。下面将对这些函数进行详细介绍。

3.1. 模式级匹配（M^{sl}）

模式级匹配利用属性名称中的信息，估算源数据库和目标数据库中任意两对属性 $\langle a^s, a^d \rangle$ 之间的匹配得分。表2分别展示了一个来自州级数据库和 NVIS 数据库的源表与目标表的示例。本文将始终采用该示例来说明不同的匹配器。模式级匹配是一种集成方法，由以下两种基于模式级的匹配器组成。

表2：州级数据库和 NVIS 数据库中的一个示例条目

(a) 源数据库（州级数据库）

u_heightCode	树种 3_
8	罗斯桉
0	桥桉
2	轮生木麻黄

(b) 目标数据库（NVIS 数据库）

u_身高类别	u_物种 3 _
0	桥桉
1	Atalaya hemiglauca
5	Pomaderris aspera

领域知识（M^{sl}_{DK}）： 第一个模式层匹配器会在可用时运用领域知识。在我们的 NVIS 场景中，我们利用领域专家积累的大量知识，这些知识记录在了

NVIS手册（NVIS技术工作组，2017）。这些规则以如果-那么- 形式编码，并输入到 SAHM 系统中。基于领域知识的模式层匹配器（记作 M_{DK}^{sl} ）遍历这些规则，以识别属性对 $\langle a^s, a^d \rangle$ 是否符合某条有效规则。也就是说，给定源数据集和目标数据集 R^s 和 R^d，以及属性对集合 $\{\langle a^s, a^d \rangle\}^{n_s \times n_d}$，针对每个属性对 $\langle a^s, a^d \rangle$ 计算基于领域知识的模式层匹配器如下：

$$M_{DK}^{sl}(\langle a^s, a^d \rangle) = \begin{cases} 1 & \text{if “if } a^s \text{ } (resp.\ a^d) \text{ then } a^d \text{ } (resp.\ a^s) \text{” is valid;} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

尽管领域知识在模式匹配中可能发挥重要作用，但它并非 SAHM 的必要条件。如果没有领域知识， M_{DK}^{sl} 可以从集成 M^{sl} 中排除，实验部分的结果表明，这样仍能取得相当不错的效果。在这种情况下，$M_{DK}^{sl}(\langle a^s, a^d \rangle) = 0$ 适用于所有配对 $\langle a^s, a^d \rangle$。

<escape>**基于语言的匹配 (M_{Lin}^{sl})：** 第二个模式层匹配器采用了 3 种基于语言的度量指标来识别匹配关系。这些指标分别是 Levenshtein 相似度（ Levenshtein et al., 1966）、Monge-Elkan 相似度（Monge et al., 1996）和词频-逆文档频率（Salton 和 McGill, 1983）。Levenshtein 相似度根据编辑距离计算两个字符串之间的相似程度。当 $a^s = a^d$ 时，Levenshtein 相似度的取值范围为 $[0, 1]$, ，其中 Levenshtein 相似度为 1。具体而言，两个字符串 a^s 和a^d 之间的 Levenshtein 相似度 Sim_{LEV}使用公式（2）计算。这里，$|a^s|$ 和 $|a^d|$ 分别表示 a^s 和 a^d中的字母数量。因此，max($|a^s|$, $|a^d|$) 表示 a^s 和 a^d中较长字符串的字母数量。</escape>

$$Sim_{\mathrm{LEV}}(a^s, a^d) = 1 - \frac{\mathrm{editDistance}(a^s, a^d)}{\max(|a^s|, |a^d|)} \tag{2}$$

例如，表2中“$u\ heightCode$”和“$u\ heightclass$”之间的 Levenshtein 距离为 5，包括在位置 8 和 12 分别插入两个字母（即“ ”、“s”），以及在“ $u\ heightCode$”的位置 10、11 和 12 分别进行三次替换（即“o” → “l”、“d” → “a”和“e” → “s”）。使用公式（2）计算得到的相似度得分为 $1 - (5/14) = 0.64$。

如图2所示，某些属性可能是由多个词组成的复杂复合名称。这是 NVIS 环境所特有的，因为属性名称可能包含植物谱系。针对这种情况，我们采用了 Monge-Elkan 和词频-逆文档频率（TFIDF）技术。在 Monge-Elkan 相似度中，针对所有词元，在词元层级计算平均 Levenshtein 相似度。
Monge-Elkan 相似度的取值范围为 $[0,1]$。给定字符串 a^s（分别为 a^d），令 T（a^s）（分别为 T（a^d））表示 a^s（分别为 a^d）中的词元。
Monge-Elkan 相似度 Sim_{ME} 按照公式（3）计算。在公式（3）中，Sim_{LEV}（T（a^s），T（a^d））表示 T（a^s）$_i$（即 a^s 中的第 i 个词元）与 T（a^d）$_j$（即 a^d 中的第 j 个词元）之间的 Levenshtein 相似度（参见 公式（2））。

$$Sim_{ME}(a^s, a^d) = \frac{1}{|T(a^s)|} \sum_{i=1}^{T(a^s)} \max\{Sim_{LEV}(T(a^s)_i, T(a^d)_j)\}_{j=1}^{T(a^d)} \quad (3)$$

例如，考虑表2中的两个属性“$treesp\ 3$”和“$u\ species\ 3$”。内部函数用于计算最佳匹配词元的分数。因此，Sim_{LEV}(‘treesp’,‘u’) $= 0$, Sim_{LEV}(‘treesp’,‘species’)$= 0.28$, Sim_{LEV}(‘treesp’,‘3’) $= 0$, Sim_{LEV}(‘3’,‘u’) $= 0$, Sim_{LEV}(‘3’,‘species’)$= 0$, Sim_{LEV}(‘3’,‘3’) $= 1$。最后，使用公式（3）计算出的 MongeElkan 相似度分数为 $1/2 * (0.28 + 1) = 0.64$。

对于词频-逆文档频率（Salton 和 McGill，1983年），源数据库和目标数据库中的每个属性名称都会被分词。接下来，将每个属性名称视为一篇文档，并将该属性中的每个词元视为一个词项。随后计算每个词元的词频-逆文档频率分数。任意两个源属性名称和目标属性名称 a^s 与 a^d 之间的词频-逆文档频率相似度分数Sim_{TFIDF}(a^s, a^d)使用公式（4）计算。

$$Sim_{TFIDF}(a^s, a^d) = \frac{\prod_{\forall a \in \{a^s, a^d\}} \{\sum_{t \in \{T(a^s) \cap T(a^d)\}} TFIDF(t, a)\}}{\prod_{\forall a \in \{a^s, a^d\}} \{\sum_{t \in \{T(a^s) \cup T(a^d)\}} TFIDF(t, a)\}} \quad (4)$$

这里，我们使用符号 $a^s \cap a^d$ 表示属性之间的共同词元，使用 $a^s \cup a^d$ 表示词元的并集。函数 $TFIDF$(t, a)用于计算给定属性 a中词元 t 的词频-逆文档频率分数。该分数

归一化至范围 $[0,1]$。例如，源属性（a^s）‘$u\ heightCode$’与目标属性（a^d）‘$u\ height\ class$’之间的 TF-IDF 相似度分数（参见 表2）计算如下：

$$Sim_{TFIDF}(u\ heightCode, u\ height\ class) = \frac{\prod_{\forall a \in \{u_heightCode, u_height_class\}} \{\sum_{t \in \{u, height\}} TFIDF(t,a)\}}{\prod_{\forall a \in \{u_heightCode, u_height_class\}} \{\sum_{t \in \{u, height, Code, class\}} TFIDF(t,a)\}}$$

$$= \frac{\left(\begin{array}{c}(TFIDF(u, u\ heightCode) + TFIDF(height, u\ heightCode)) \times (TFIDF \\ (u, u\ height\ class) + TFIDF(height, u\ height\ class))\end{array}\right)}{\left(\begin{array}{l}(TFIDF(u, u_heightCode) + TFIDF(height, u_heightCode) \\ \quad + TFIDF(Code, u_heightCode) + TFIDF(class, u_heightCode)) \\ \times \quad (TFIDF(u, u_height_class) + TFIDF(height, u_height_class) \\ \quad + TFIDF(class, u_height_class) + TFIDF(Code, u_height_class))\end{array}\right)}$$

$$= \frac{(0.04 + 0.10) \times (0.04 + 0.10)}{(0.04 + 0.10 + 0.20 + 0) \times (0.04 + 0.10 + 0.20 + 0)} = \frac{0.0196}{0.1156} = 0.17$$

同样，TF-IDF 相似度分数 $Sim_{TFIDF}(u\ heightCode, u\ species3)$（*cf.*表2）为 0.013，低于 $Sim_{TFIDF}(u\ heightCode, uheight\ class) = 0.17$ 的分数，这是因为前者包含更多不相似的词元。

计算各个语言学分数后，通过加权和生成任意一对 $\langle a^s, a^d \rangle$ 的最终语言学相似度分数。也就是说，给定源数据集和目标数据集 R^s 和 R^d，以及属性对集合 $\{\langle a^s, a^d \rangle\}^{n_s \times n_d}$，每一对 $\langle a^s, a^d \rangle$ 的语言学模式级匹配器计算如下：

$$M^{sl}_{Lin}(\langle a^s, a^d \rangle) = g_1 \cdot Sim_{Lev}(\langle a^s, a^d \rangle) + g_2 \cdot Sim_{ME}(\langle a^s, a^d \rangle) + g_3 \cdot Sim_{TFIDF}(\langle a^s, a^d \rangle) \quad (5)$$

其中 $g_1 + g_2 + g_3 = 1$。

需要注意的是，所使用的语言学度量指标的数量和类型可以视为用户自定义参数。然而，我们的实验表明，对于我们的应用场景，当前配置能够获得相对更准确的结果。

3.2. 实例级匹配（M^{il}）

实例级匹配利用源数据库和目标数据库中任意两个属性的属性值信息 $\langle a^s, a^d \rangle$ 来估算它们之间的匹配分数。该集成由以下 2 个基于实例级匹配的匹配器组成。

单变量统计匹配 (M^{il}_{Uni})： 该匹配器利用属性值计算一对属性之间的相似度 $\langle a^s, a^d \rangle$ 。该匹配器有两种变体：一种用于类别型属性，另一种用于数值型属性。对于数值型属性，我们使用属性的均值和标准差；而对于类别型属性，我们使用 Jaccard 系数。具体而言，对于数值型属性，均值用于确定两个属性的中心趋势，并据此衡量二者的接近程度。此外，还可以利用它们的标准差来衡量各自离散程度与相应中心趋势的接近程度。我们采用均值和标准差的归一化绝对差值的加权和来计算单变量得分。对于两个类别型属性，我们采用 Jaccard 系数（Chaudhuri 等，2006）。Jaccard 系数确保两个属性共有的属性值越多，它们就被认为越相似。因此，给定源数据集和目标数据集 R^s 和 R^d，以及属性对集合 $\{\langle a^s, a^d \rangle\}^{n_s \times n_d}$，每个属性对 $\langle a^s, a^d \rangle$ 的单变量统计实例级匹配器计算如下：

$$M^{il}_{Uni}(\langle a^s, a^d \rangle) = \begin{cases} 1 - \left(\frac{0.8 \cdot |\mu(a^s) - \mu(a^d)|}{\max(\mu(a^s), \mu(a^d))} + \frac{0.2 \cdot |SD(a^s) - SD(a^d)|}{\max(SD(a^s), SD(a^d))}\right) & \text{if } a^s, a^d \\ & \text{are numerical} \\ \frac{|\{T(a^s) \cap T(a^d)\}|}{|\{T(a^s) \cup T(a^d)\}|} & \text{otherwise} \end{cases} \tag{6}$$

在公式 (6) 中，μ 和 SD 分别是均值和标准差。$M^{il}_{Uni}(\langle a^s, a^d \rangle) \in [0, 1]$。请注意，当属性为数值型且其均值和标准差彼此接近时，得分趋近于 $1\ i.e\ M^{il}_{Uni}(\langle a^s, a^d \rangle) \rightarrow 1$。类似地，当属性为类别型且共有属性值的数量彼此接近时，得分也趋近于 $1\ i.e\ M^{il}_{Uni}(\langle a^s, a^d \rangle) \rightarrow 1$。我们在公式 6 中为均值赋予了更大的权重，因为它在其中更为重要。

根据我们的实验结果，计算相似度贡献值。如果两个属性中一个为数值型，另一个为类别型，则首先通过离散化将数值型属性转换为类别型属性（Rahman & Islam, 2016），然后计算 Jaccard 系数。

例如，在表2中，既包含数值型属性，也包含类别型属性。因此，属性对 $\langle u\ heightCode, u\ height\ class\rangle$ 的单变量得分为，

$$
\begin{aligned}
&M_{Uni}^{il}(\langle u_heightCode, u_height_class\rangle) \\
&\quad = 1 - \Big(\frac{0.8 \cdot |\mu(u_heightCode) - \mu(u_height_class)|}{\max(\mu(u_heightCode), \mu(u_height_class))} \\
&\quad + \frac{0.2 \cdot |SD(u_heightCode) - SD(u_height_class)|}{\max(SD(u_heightCode), SD(u_height_class))}\Big) \\
&\quad = 1 - \Big(\frac{0.8 \cdot |3.33 - 2|}{\max(3.33, 2)} + \frac{0.2 \cdot |3.39 - 2.16|}{\max(3.39, 2.16)}\Big) \\
&\quad = 0.61
\end{aligned}
$$

再次，属性对 $\langle treesp3, u\ species3\rangle$ 的单变量得分为

$$
\begin{aligned}
&M_{Uni}^{il}(\langle treesp3, u_species3\rangle) \\
&\quad = \frac{|\{T(treesp3) \cap T(u_species3)\}|}{|\{T(treesp3) \cup T(u_species3)\}|} \\
&\quad = \frac{1}{5} \\
&\quad = 0.2
\end{aligned}
$$

多变量统计匹配 (M_{Mul}^{il})： 这种新型匹配器基于源属性对与目标属性对之间的相关性 $\langle a^s, a^d\rangle$。其基本思路是：对于已知匹配的一对属性 $\langle a^s, a^d\rangle$，源数据库 R^s 中 a^s 与其他属性之间的相关性，与目标数据库 R^d中 a^d 与其他属性之间的相关性具有相似特征。对于已知匹配的给定属性对 $\langle a^s, a^d\rangle$，多变量统计匹配得分 M_{Mul}^{il} 会计算源属性 a^s 与源数据集 R^S中其他属性 a_j^s 之间的皮尔逊相关系数。随后，根据皮尔逊相关系数得分生成属性的排序列表 $[(a_q^s, {}^{pc_{a_q^s}}) \preceq \cdots \preceq (a_r^s, {}^{pc_{a_r^s}})]$。同样，多变量统计匹配得分 M_{Mul}^{il} 会计算目标属性 a^d 与目标数据集 R^d中其他属性 a_k^d之间的皮尔逊相关系数。随后也会根据皮尔逊相关系数得分生成属性的排序列表 $[(a_q^d, {}^{pc_{a_q^d}}) \preceq \cdots \preceq (a_r^d, {}^{pc_{a_r^d}})]$。对于每一组属性两两组合

$[(a_q^s, pc_{a_q^s}) \preceq \cdots \preceq (a_r^s, pc_{a_r^s})]$ 中的属性与 $[(a_q^d, pc_{a_q^d}) \preceq \cdots \preceq (a_r^d, pc_{a_r^d})]$, 中的属性两两组合，即对于每一对 $< a_j^s, a_k^d >$，都会计算其皮尔逊相关系数值的绝对距离 $|pc_{a_j^s} - pc_{a_k^d}|$。因此，对于任意一对已知匹配的属性 $\langle a^s, a^d \rangle$，都会生成一组距离值 $\{|pc_{a_j^s} - pc_{a_k^d}|\}^{(n_d-1)\times(n_s-1)}$，其中 n_s（分别为 n_d）是源数据库（分别为目标数据库）中的属性数量。将距离值除以最大距离 2 进行归一化，再用 1 减去归一化后的值进行反转。因此，对于给定的 $\langle a^s, a^d \rangle$，$M_{Mul}^{il}(\langle a^s, a^d \rangle)$会生成一组相似度分数 $M_{Mul}^{il}(\langle a^s, a^d \rangle) = \{Sim_{pc}(a_j^s, a_k^d)\}^{(n_d-1)\times(n_s-1)}$，其中 $Sim_{pc}(a_j^s, a_k^d) = 1 - \frac{|pc_{a_j^s} - pc_{a_k^d}|}{2}$。由此，对于所有先前已知的匹配对$\{\langle a^s, a^d \rangle\}$，可以为每一对 $\langle a^s, a^d \rangle$ 计算 $M_{Mul}^{il}(\langle a^s, a^d \rangle)$，并对每个先前已知匹配对 $\langle a^s, a^d \rangle$生成的相应分数 $Sim_{pc}(a_j^s, a_k^d)$取平均，得到最终得分 $\bar{Sim}_{pc}(a_j^s, a_k^d)$。

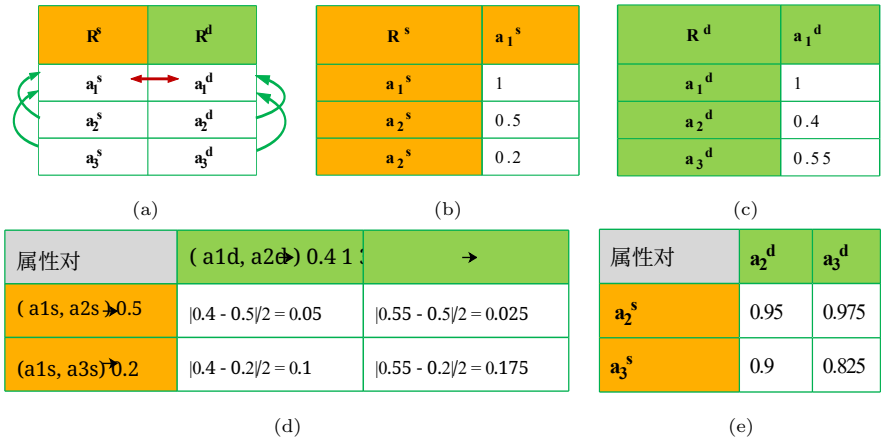


图 3：多变量归一化评分示例。(a) 源属性与目标属性列表，其中用户定义的属性对为 $(a_1^s \leftrightarrow a_1^d)$；(b) 包含相关性得分的源表；(c) 包含相关性得分的目标表；(d) 源-目标属性对的距离计算；(e) 多变量匹配得分

图 3 展示了一个多变量评分方案的示例场景，其中 $\{a_1^s, a_2^s, a_3^s\}$ 和 $\{a_1^d, a_2^d, a_3^d\}$ 分别是源数据集和目标数据集中的属性（图 3a）。在所提出的归一化多变量评分方案中，首先需要用户输入，由用户为特定的源数据集与目标数据集配对定义一个已知匹配关系。在图 3a 的示例中，这一由用户定义的配对（a_1^s, a_1^d）用红色箭头表示。由于其中一对之间的关系已经定义，因此所提出的多变量.....

多变量评分方案将为另外两个源属性推荐目标属性。源属性之间的多变量得分如图 3b 所示，该图根据用户定义的关系展示了源属性对之间的关联程度。数值越高，表示关联越强。类似地，图 3c 展示了目标属性对之间的多变量得分。对于每一种源属性与目标属性的组合，图 3d 展示了从源属性表和目标属性表中获得的多变量得分的归一化绝对距离值。由于图 3d 展示的是源属性与目标属性之间的归一化绝对距离，因此得分越低，表示两者的相关性越强。最后，通过用 1 减去上一步得到的各项得分，计算得到多变量得分（图 3e）。

如算法 1 所示，使用加权和 $w_1 \cdot M_1^{sl} + w_2 \cdot M_{Lin}^{sl} + w_3 \cdot M_{Uni}^{il} + w_4 \cdot M_{Mul}^{il}$来生成每对源属性和目标属性的最终匹配分数。其中 $w_1 + w_2 + w_3 + w_4 = 1$。在此方案中，我们为各权重赋予了相同的值；不过，根据问题的具体情境，可以对 SAHM 中的不同匹配器赋予不同的优先级。

本节介绍了一种基于数据库模式和实例的新型两阶段模式匹配技术。下一节将展示我们提出的新方法 SAHM 的有效性。

4. 实验与结果

本节将对所提出的方法 SAHM 进行性能评估，并与 Valentine 框架中列出的现有知名模式匹配与映射技术进行比较（Koutras 等，2021）。这些技术包括 COMA、CUPID、相似性泛洪、Jaccard-Leven匹配器、基于分布的匹配器和 EmbDI。在评估中，我们考察以 F1 分数和准确率衡量的有效性，以及以运行时间衡量的效率。此外，我们还进行了消融研究，以探究所提出方法 SAHM 中各匹配器的贡献。

算法 1： SAHM 评分算法

算法 `SAHM()`

对 $a^d \longleftarrow R^d$ **执行 foreach**

  对 $a^s \longleftarrow R^s$ **执行 foreach**

    $score^{sl}_{DK}, score^{sl}_{Lin} \longleftarrow$ SLM(a^s, a^d)

    $score^{il}_{Uni}, score^{il}_{Mul} \longleftarrow$ ILM(a^s, a^d)

    $score(a^s, a^d) \longleftarrow$

      $w_1 \cdot score^{sl}_{DK} + w_2 \cdot score^{sl}_{Lin} + w_3 \cdot score^{il}_{Uni} + w_4 \cdot score^{il}_{Mul}$

return $score(a^s, a^d)$

过程 `SLM(`a^s, a^d`)`**：** $M^{sl}_{DK}, M^{sl}_{Lin}$

计算 M^{sl}_{DK} 使用公式（1）

计算 M^{sl}_{Lin} 使用公式（5）

return $M^{sl}_{DK}, M^{il}_{Lin}$

过程 `ILM(`a^s, a^d`)`**：** $M^{il}_{Uni}, M^{il}_{Mul}$

计算 M^{il}_{Uni} 使用公式（6）

对 $\langle a^s, a^d \rangle \longleftarrow Known\ Matching\ Pairs\ \{\langle a^s, a^d \rangle\}$ **执行 foreach**

  $[(a^s_q, pc_{a^s_q}) \preceq \cdots \preceq (a^s_r, pc_{a^s_r})] \longleftarrow$ CalcCorrelation(a^s, a^s_j)

  $[(a^d_q, pc_{a^d_q}) \preceq \cdots \preceq (a^d_r, pc_{a^d_r})] \longleftarrow$ 计算相关性(a^d, a^d_k)

  $CR_v \longleftarrow$ 计算镜像相关性($[(a^s_q, pc_{a^s_q}) \preceq \cdots \preceq (a^s_r, pc_{a^s_r})], [(a^d_q, pc_{a^d_q}) \preceq \cdots \preceq (a^d_r, pc_{a^d_r})]$)

  添加 CR_v 进入 M^{il}_{Mul}

return $M^{il}_{Uni}, M^{il}_{Mul}$

过程 `CalcCorrelation (`a^*, a^*_*`)`**：** CR_*

$CR_* = []$

foreach $a \longleftarrow a^*_*$ **do**

  $\langle a^*_q, pc_{a^*_q} \rangle \longleftarrow$ 皮尔逊相关系数(a^*, a)

  添加 $\langle a^*_q, pc_{a^*_q} \rangle$ 至 CR_*

return CR_*

过程 `CalcMirrorCorrelation(`CR_s,CR_d`)`**：** CR_L

对每个 a^s_j，$pc_{a^s_j} \longleftarrow CR_s$ **执行**

  foreach $a^d_k, pc_{a^d_k} \longleftarrow CR_d$ **do**

    $Dist \longleftarrow \|pc_{a^s_j} - pc_{a^d_k}\|$

    $Sim_{pc}(a^s_j, a^d_k) \longleftarrow 1 -$ (Dist/2)

    if $Sim_{pc}(a^s_j, a^d_k)$ $not\ exist\ in\ CR_L$ **则**

      在 CR_L 中添加 $Sim_{pc}(a^s_j, a^d_k)$

    else

      用现有 $Sim_{pc}(a^s_j, a^d_k)$ 的平均值更新 CR_L 和新值 $Sim_{pc_{(a^s_j, a^d_k)}}$

return CR_L

表 3：专有 NVIS 数据集

数据集	# 属性	# 记录
ACT	35	19226
NT	40	588
WA	17	2158
SA	20	1167
QLD	20	14610
NVIS	68	10775

表 4：音乐家数据集

数据集	#属性	#记录	描述
可联合的	20	10846	A 源与目标的配对 具有相同属性集的数据集- 可能是原文照录或ышәтә 含有噪声。它还包含一些 com- 可能是动词性的实例，- 时间或含噪声
View 可联合的	13	5423	A 源和目标的一对 数据集之间没有共同的英- 例子，但包含一些常见的 可能逐字复制的属性 或经过噪声处理。
可连接性	13	10846	A 源与目标的配对 具有一些共同特征的数据集 实例以及一些常见的 ат- 属性。 共同属性 可能是逐字原文，也可能经过了噪声处理。
语义上 可连接的	14	10846	A 源数据集与目标数据集的 数据集，它们具有一些共同的 实例和一些共同的属- 性。 共同实例 且属性可能是原文照录的，也可能是 经过加噪 处理的。

实验设置：所有实验均在 Windows 10 操作系统上进行，使用配备 AMD Ryzen 7 (PRO 3700Uw) 2.30GHz 处理器和 32 GB 内存的计算机。所提出方法的性能在 NVIS 数据集和音乐家数据集上均进行了评估。NVIS 数据集包含来自西澳大利亚州（WA）、南澳大利亚州（SA）、昆士兰州（QLD）、北领地（NT）和澳大利亚首都领地（ACT）的专有州数据集。表 3 展示了这些数据集的详细信息。每个州数据集均作为源数据集，与 NVIS 数据集（目标数据集）进行匹配。

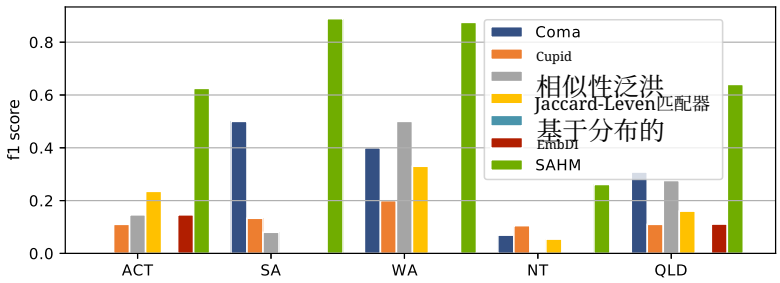
音乐家数据集 [3] 是由（Koutras 等人，2021）从 WikiData [4]生成的公开可用的美国音乐家数据集。该音乐家数据集有 4 个变体，详见表 4。

在我们所有的实验中，评估均基于源数据集与目标数据集之间的匹配效果。值得注意的是，我们提出的方法 SAHM 的输出结果是：针对目标数据集中的每个属性，对源数据集中的属性进行排名。为了计算准确率，我们考虑排名靠前的 N 个属性。我们用 *Top ** 表示排名靠前的 N 个属性，其中 * 是用于确定所考虑属性数量的参数。默认情况下，我们始终使用 *Top 1*，但用户可以更改此参数，以获得更多排名靠前的建议。例如，在图 6 中，我们展示了根据 SAHM 得出的排名前 1 至 4 的属性（*i.e. Top-1, Top 2, Top 3 & Top 4*）的使用效果。此外，在所有实验中，我们均不考虑 SAHM 的领域知识匹配器，因为显而易见，领域知识匹配器能够与真实匹配结果保持一致。

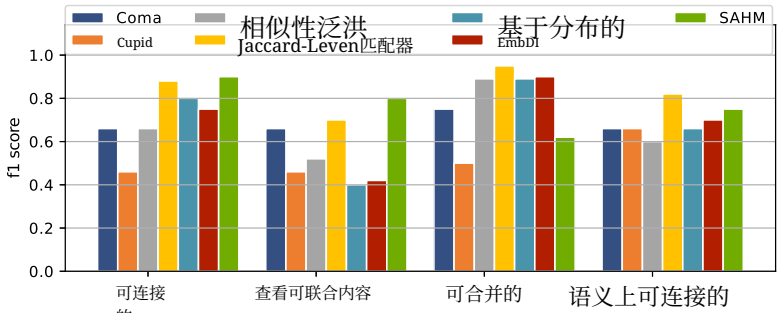
4.1. SAHM 的有效性

在本实验中，我们通过计算 F1分数，将 SAHM 的有效性与另外五种最先进的模式匹配算法进行了比较。在此比较中，对于 COMA、CUPID、相似性泛洪、Jaccard-Leven匹配器、基于分布的匹配器和 EmbDI，我们采用了 Valentine 框架的标准实验设置，以找出州数据集与 NVIS 数据集之间可能的最佳匹配。在 SAHM 中，我们将 *Top 1* 视为最佳匹配。

[3]可在 https://bit.ly/3Kk3ryU[4]
https://www.wikidata.org/wiki/Wikidata 获取



(a)



(b)

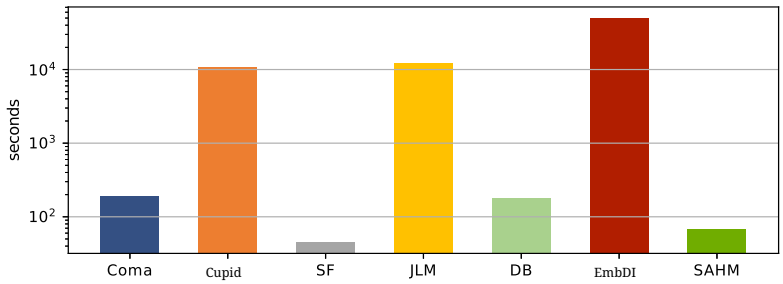
图 4：不同模式匹配算法的 F1 分数比较（a）专有 NVIS 数据集（b）音乐家数据集

图 4a 展示了我们在专有数据集上进行实验的结果。可以看出，在所有数据集上，我们提出的方法总体 F1分数均优于其他方法。总体而言，由于属性的命名模式较为特殊，北领地州数据集上的表现低于其他州。值得注意的是，结果表明，现有方法并非总是适用。也就是说，现有技术预测的最佳匹配并不总是与真实情况一致。例如，可以看出，COMA 无法在 ACT 数据集与 NVIS 数据集之间正确识别任何真实匹配，而 Jaccard-Leven匹配器在南澳大利亚州数据集中未能找到任何匹配。此外，EmbDI 无法在南澳大利亚州、西澳大利亚州和北领地数据集中识别出任何匹配。基于分布的方法表现最差，因为它在所有数据集中都未能找到任何真实匹配。这些结果不仅展示了 SAHM 的有效性，也表明其在不同州数据集上的可靠性。

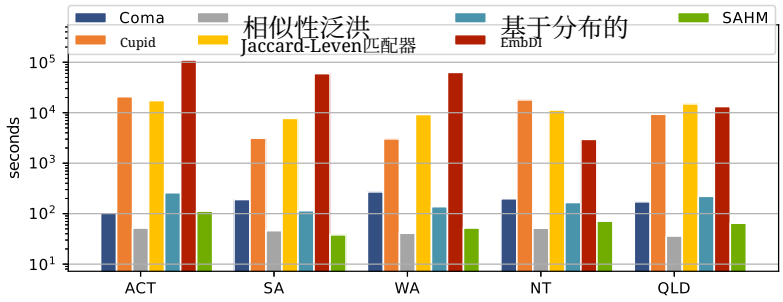
图 4b 展示了从公开可用的音乐家数据集获得的结果。我们观察到，SAHM 的竞争力很强，并且在四个数据集中有两个表现最佳。特别是，我们注意到，SAHM 在属性数量较少的音乐家数据集上表现最好。在包含 13 个属性的可连接数据集和视图可并集数据集中，SAHM 的表现最佳；在包含 14 个属性的语义可连接数据集中，表现排名第二。然而，在包含 20 个属性的可并集数据集中，SAHM 的竞争力不足。回顾表 4 可知，可并集数据集中的源数据集和目标数据集具有相同的属性集合，但源数据集中的属性经过了加噪处理。在这种情况下，SAHM 无法有效区分这些属性。这是由噪声的生成方式导致的。噪声通常会在源数据集的属性后添加“Label”（*e.g.* “cityLabel”）。在这种情况下，模式层匹配器通常会将这些属性与目标数据集中的属性错误匹配（*e.g.* “religionLabel”）。我们观察到，在这种情况下，如果调整 SAHM 的加权方案，使其更偏向基于实例的匹配组成部分，SAHM 的表现会显著提升。未来，我们将探讨如何针对任意给定数据集优化这一加权方式。

4.2. SAHM 的效率

在本实验中，我们使用专有的 NVIS 数据集计算运行时间，并将 SAHM 的效率与其他五种模式匹配算法进行了比较。图 5 展示了效率实验结果，采用对数刻度。图 5a 显示了各算法在所有数据集上的平均运行时间。我们观察到，SAHM 的效率比大多数现有先进技术高出几个数量级。只有 Similirity Flooding（SF）的效率略高。进一步分析图 5b 中各数据集的单独运行时间结果可以发现，对于每个州数据集，SAHM 和 SF 都是效率最高的算法。与 SAHM 相比，SF 的效率更高，因为 SF 是纯粹基于模式的匹配器，而 SAHM 则结合了模式匹配和实例级匹配。我们还观察到，EmbDI 耗时最长，这是因为其随机游走生成存在瓶颈；随着实例数量增加，其扩展能力不佳。此外，EmbDI 依赖词嵌入，这也增加了其计算成本。值得注意的是，尽管 SF 是唯一一个效率略高于 SAHM 的匹配算法，但这种效率优势需要付出相当大的代价。



(a)



(b)

图 5：效率分析。(a) 各匹配算法处理所有州数据集的平均运行时间 (b) 各州及相应匹配算法的运行时间结果。

效果方面。也就是说，如图 4 所示，旧金山的效果并不具有竞争力。

4.3. SAHM 根据推荐数量的性能

在本次实验中，我们旨在展示参数 *Top ** 对 SAHM 准确性的影响。结果如图 6 所示。可以观察到，对于所有数据集， *Top 1* 的表现都令人满意，但随着参数 *Top ** 增加至 *Top 4*，其表现大幅提升。图 6 中的分析表明，四个州数据集 *i.e.* 澳大利亚首都领地、昆士兰州、南澳大利亚州和西澳大利亚州的前四项建议准确率介于 88% 至 100 % 之间。然而，¯北领地数据集的准确率最高仅为 64%。与此前观察到的一样，北领地数据集的表现最差，因为其属性名称采用了不常见的命名规范。

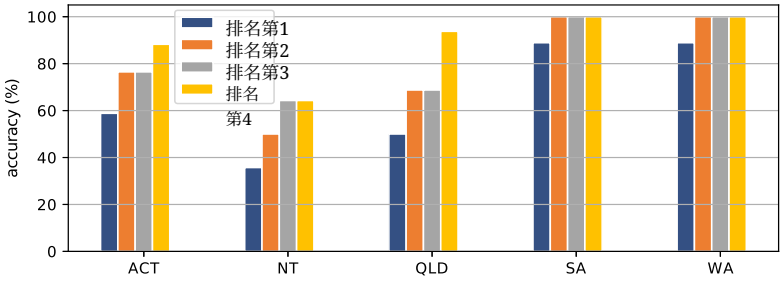


图 6：基于模式匹配建议的州数据集准确率百分比。

表 5：建议数量对各组成部分准确率的影响

单个组成部分			西澳大利亚州（%）	南澳大利亚州（%）	昆士兰州(%)	北领地(%)	澳大利亚首都领地(%)
语言相似性组成部分		排名第1	66.66	77.77	37.5	14.29	35.29
		排名第2	66.66	88.88	50	57.15	70.58
		排名第3	66.66	99.99	62.5	57.15	76.46
		排名第4	66.66	99.99	75	57.15	82.34
单变量组成部分		排名第1	55.55	44.44	25	21.43	35.29
		排名第2	66.66	44.44	43.75	21.43	64.7
		排名第3	66.66	44.44	56.25	21.43	76.46
		排名第4	77.77	55.55	56.25	28.57	76.46
Multivariate Component	随机匹配 Pair	排名第1	55.55	22.22	18.75	7.14	0
		排名第2	88.88	22.22	25	14.28	0
		排名第3	88.88	33.33	37.5	28.57	5.88
		排名第4	88.88	33.33	43.75	35.71	17.64
	已知匹配项 Pair	排名第1	66.66	22.22	18.75	21.43	41.17
		排名第2	88.88	44.44	31.25	42.86	47.05
		排名第3	88.88	55.55	37.5	57.15	52.93
		排名第4	88.88	66.66	50	64.29	58.81

4.4. SAHM 各组成部分的性能

在本实验中，我们旨在展示 SAHM 各个组成部分的有效性，以及推荐数量对各个匹配器的影响。结果如表 5 所示。我们提醒读者，在多变量统计匹配中，使用源属性与目标属性之间已知的真实匹配对作为计算其他匹配结果的基础。在表中，多变量（随机 .....

随机匹配对（随机匹配对）表示用户没有任何真实的目标属性与源属性匹配对，因此随机选择一对。多变量（已知匹配对）表示用户拥有一对真实的目标属性与源属性匹配对。

该表展示了语言相似性匹配、单变量统计匹配和多变量统计匹配的结果。请注意，由于领域知识匹配的结果显而易见，因此未采用该方法。从表中可以看出，在所有情况下，随着推荐数量的增加（$i.e.\ Top\ *$），准确率均呈单调递增趋势。有趣的是，WA 数据集的语言相似性 $Top\ 1$ 为 66.66%，且该数值不会随着推荐数量的增加而改变。这一数值表明，如果我们仅考虑框架中的语言相似性组成部分来处理 WA 数据集，那么该数据集的模式匹配准确率将始终为 66.66%。显然，对于 WA 数据集，多变量组成部分的性能最佳。此外，我们观察到，对于 SA 数据集，语言相似性是最重要的组成部分。在 QLD 和 ACT 数据集中，也可以观察到语言相似性具有较强的性能。然而，在 NT 数据集中，多变量匹配器（已知匹配对）取得的结果远优于语言相似性匹配。这表明，SAHM 能够针对不同数据集充分利用不同匹配器的适用性，从而受益于集成方法。

5. 结论

本研究开发了一种名为 SAHM 的半自动数据库模式匹配框架，用于匹配来自异构数据源的数据。SAHM 是一个双层框架，由模式级匹配和实例级匹配组成。每一层都整合了多种技术，这些技术协同提升了匹配框架的有效性。此外，本研究提出了一种新颖的多变量统计匹配技术，并证明该技术能够增强现有的模式匹配技术。研究使用专有的 NVIS 数据集对 SAHM 进行了评估，结果表明其性能远优于现有的最先进技术。研究结果显示，通过在新颖的 SAHM 框架中利用领域知识，模式匹配的有效性能够大幅提升。同时，在可能缺乏领域知识的情形下，SAHM 仍能取得合理的表现。

以获得更好的结果。未来，我们计划研究所提出的模式匹配框架如何提高其在下游机器学习应用中的实用性。此外，还将研究其他相似性度量方法，例如语义相似性和基于嵌入的相似性方法。

致谢

本研究由澳大利亚气候变化、能源、环境与水务部（DCCEEW）（原农业、水务与环境部（DAWE））以及澳大利亚 CSIRO Data61 资助。

参考文献

Aumueller, D., Do, H.-H., Massmann, S., & Rahm, E. (2005). 使用 Coma++ 进行模式与本体匹配。在 *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (第 906–908 页)。

Bilke, A., & Naumann, F. (2005). 利用重复项进行模式匹配。载于 *21st International Conference on Data Engineering (ICDE'05)*（第 69–80 页）。IEEE。Cappuzzo, R., Papotti, P., & Thirumuruganathan, S. (2021). Embdi：为关系数据集成生成嵌入。载于 *SEBD 2021: The 29th Italian Symposium on Advanced Database Systems*。Chaudhuri, S., Ganti, V., & Kaushik, R. (2006). 数据清洗中相似性连接的一种基本算子。载于 *22nd International Conference on Data Engineering(ICDE'06)*（第 5–5 页）。IEEE。Chen, C., Golshan, B., Halevy, A. Y., Tan, W.-C., & Doan, A. (2018). Biggorilla：一个用于数据准备与集成的开源生态系统。 *IEEE Data Eng.Bull.*, *41*, 10–22。Coen, G., & Xue, P. (2009). 利用数据字典进行模式匹配。载于 *International Conference on Application of Natural Language to Information Systems*（第 62–78 页）。Springer。Cruz, I. F., Antonelli, F. P., & Stroe, C. (2009). Agreementmaker：面向大型真实世界模式和本体的高效匹配。 *Proceedings of the VLDB Endow-ment*, *2*, 1586–1589。

Do, H.-H., & Rahm, E. (2002). Coma——一种用于灵活组合模式匹配方法的系统。In *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases* (pp. 610–621). 爱思唯尔。Do, H.-H., & Rahm, E. (2007). 大型模式匹配：方法与评估。*Information Systems*, *32*, 857–885。Dongming, Z., & Guohua, L. (2010). Esm：一种数据库模式之间的扩展模式匹配方法。In *2010 International Conference on E-Business andE-Government* (pp. 3996–4000). IEEE。Fernandez, R. C., Mansour, E., Qahtan, A. A., Elmagarmid, A., Ilyas, I., Madden, S., Ouzzani, M., Stonebraker, M., & Tang, N. (2018). 语义渗透：利用词嵌入连接数据集以实现数据发现。In *2018 IEEE 34th International Conference on Data Engineering (ICDE)* (pp. 989–1000). IEEE。Gal, A., Roitman, H., & Shraga, R. (2019). 学习对模式匹配结果进行重新排序。 *IEEE Transactions on Knowledge and Data Engineering*, 。Haslem, A., Callister, K. E., Avitabile, S. C., Griffioen, P. A., Kelly, L. T., Nimmo, D. G., Spence-Bailey, L. M., Taylor, R. S., Watson, S. J., Brown, L. et al. (2010). 一种在广阔空间范围内进行植被制图的框架：一项有助于跨行政管辖边界开展土地管理的技术。 *Landscape and Urban Planning*, *97*, 296–305。Hu, W., Qu, Y., & Cheng, G. (2008). 大型本体匹配：一种分而治之方法。 *Data & Knowledge Engineering*, *67*, 140–160。Kang, J., & Naughton, J. F. (2003). 基于不透明列名和数据值的模式匹配。In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data* (pp. 205–216)。Klein, C. J., Wilson, K. A., Watts, M., Stein, J., Carwardine, J., Mackey, B., & Poss- ingham, H. P. (2009). 纳入荒野质量的空间保护优先级排序：澳大利亚生物多样性案例研究。 *Biological Conservation*, *142*, 1282–1290。Koutras, C., Siachamis, G., Ionescu, A., Psarakis, K., Brons, J., Fragkoulis, M., Lofi, C., Bonifati, A., & Katsifodimos, A. (2021). Valentine：评估用于数据集发现的匹配技术。In *2021 IEEE 37th International Conference on Data Engineering (ICDE)* (pp. 468–479). IEEE。

Levenshtein, V. I. 等（1966）。能够纠正删除、插入和反转的二进制码。载于 *Soviet physics doklady*（第 707–710 页）。苏联，第 10 卷。Madhavan, J., Bernstein, P. A. 和 Rahm, E.（2001）。使用 Cupid 的通用模式匹配。载于 *vldb*（第 49–58 页）。Citeseer，第 1 卷。Melnik, S., Garcia-Molina, H. 和 Rahm, E.（2002）。相似性泛洪：一种通用的图匹配算法及其在模式匹配中的应用。载于 *Proceedings 18th International Conference on Data Engineering*（第 117–128 页）。IEEE。Monge, A. E., Elkan, C. 等（1996）。字段匹配问题：算法与应用。载于 *Kdd*（第 267–270 页）。第 2 卷。NVIS技术工作组（2017）。澳大利亚植被属性手册：国家植被信息系统，第 7.0 版。 *Department of the Environmentand Energy, Canberra*，。Rahm, E. 和 Bernstein, P. A.（2001）。自动模式匹配方法综述。 *the VLDB Journal*, *10*, 334–350。Rahman, M. G. 和 Islam, M. Z.（2016）。通过低频数值和属性相互依赖关系对连续属性进行离散化。 *Expert Systemswith Applications*, *45*, 410–423。Salton, G. 和 McGill, M. J.（1983）。 *Introduction to modern information retrieval*。麦格劳-希尔。Seligman, L., Mork, P., Halevy, A., Smith, K., Carey, M. J., Chen, K., Wolf, C., Madhavan, J., Kannan, A. 和 Burdick, D.（2010）。Openii：一个开源信息集成工具包。载于 *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*（第 1057–1060 页）。Shrestha, M., Tran, T. X., Bhattarai, B., Pusey, M. L. 和 Aygun, R. S.（2019）。蛋白质结晶筛选板中采用一致命名的模式匹配与数据集成。

IEEE/ACM transactions on computational biology and bioinformatics, *17*, 2074–2085。Smith, K., Morse, M., Mork, P., Li, M., Rosenthal, A., Allen, D., Seligman 和 Wolf, C.（2009）。模式匹配在大型企业中的作用。 *arXiv preprintarXiv:0909.1771*，。

Thackway, R., Lee, A., Donohue, R., Keenan, R. J. 和 Wood, M.（2007）。澳大利亚的植被信息有助于改进自然资源管理。 *Landscapeand Urban Planning*, *79*, 127–136。Zhang, M., Hadjieleftheriou, M., Ooi, B. C., Procopiuc, C. M. 和 Srivastava, D.（2010）。关于多列外键发现。 *Proceedings of the VLDB Endowment*, *3*, 805–814。Zhang, M., Hadjieleftheriou, M., Ooi, B. C., Procopiuc, C. M. 和 Srivastava, D.（2011a）。关系数据库中属性的自动发现。载于 *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*（第109–120 页）。Zhang, M., Hadjieleftheriou, M., Ooi, B. C., Procopiuc, C. M. 和 Srivastava, D.（2011b）。关系数据库中属性的自动发现。载于 *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*（第109–120 页）。